Correlations Of Categorical Data And Random Matrix Theory

A thesis submitted towards partial fulfilment of
BS-MS Dual Degree Programme

by

Aashay Patil

under the guidance of

Dr. M. S. Santhanam

Indian Institute of Science Education and Research Pune

Certificate

This is to certify that this thesis entitled "Correlations of Categorical Data and Random Matrix Theory" submitted towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research Pune represents original research carried out by Aashay Patil at IISER Pune, under the supervision of Dr. M. S. Santhanam during the academic year 2013-2014.

Student
Aashay Patil

Supervisor
Dr.M.S.Santhanam

Acknowledgements

I am highly indebted to my project supervisor, Dr. M. S. Santhanam, who I think is one of the coolest faculty in IISER Pune, for all his guidance and support throughout the year. He gave me a lot of freedom, always let me work at my own pace, never giving me any deadlines. In addition to academics, he always supported me to pursue my other interests, including granting me a 3-week leave to visit Dantewada. I learnt a lot from him, about research, life, everything.

Apart from him, there were plenty of people who made my stay at IISER memorable. I would start by expressing my sincere gratitude towards my awesome friends Bhavesh and Haritej, who is also my roommate, for all the support, discussions-academic and otherwise and constant motivation. I would like to thank my close friend Aditi, for standing besides me and cheering me up during my difficult times. My sincere thanks to my great friends Neha Bora and Kshiti for the refreshing discussions over a cup of tea. I am indebted to Aravind, Adwiteey and Rohit (Chika) for all the help with programming-related issues, to Mihir, Varun and Shibananda for helping me with Latex. Also thanks to my friends Avani, Akash G, Punya, Anurag Agrawal (Mota Bhai), Siddhartha Das (Kaka), Krishna, Vikash (Papa), Vimlesh, Indra (mama) and others whom I have missed for being awesome and giving me a plethora of memories to cherish. Last but not the least, I would like to thank my parents, for supporting me through all times.

And yes, thanks also to DST for the INSPIRE Fellowship. Life would have been a lot more difficult without the INR 5000 deposited in my account every month.

Abstract

The Random Matrix Theory (RMT) has been of great interest to physicists, with its applications in understanding the statistical structure of empirical correlation matrices appearing in the study of various real systems. However, many real systems consist of data that is not in direct numerical form. For detailed statistical analysis using RMT, converting the non-numerical data, also known as categorical data, into numerical, non-categorical data is absolutely essential. We propose a novel way to calculate correlation among objects that are not numbers. We then study various statistical properties of such random correlation matrices, such as eigenvalue density, eigenvector distribution and spacing distribution and compare and contrast them with the known results from RMT. The usefulness of such correlation calculations is demonstrated in real life applications like election result analysis and statistics of atmospheric pressure data.

Contents

Chapter 1

Introduction

Many important properties of physical systems can be represented mathematically as matrix problems. The study of statistical properties of matrices with independent and identically distributed (iid) random elements, also known as random matrices, can provide a wealth of information about the physical system it represents. Random Matrix Theory (RMT) was developed by Wigner, Dyson, Mehta and others in order to explain the energy levels of complex nuclei in Nuclear Physics [1]. They postulated that the spacings between the lines in the spectrum of a heavy atom should resemble the spacings between the eigenvalues of a random matrix, and should depend only on the symmetry class of the underlying Hamiltonian [2]. Since then, there have been rapid developments in RMT. It has found numerous applications in varied areas like Finance, Risk Management, Meteorological studies and so on.

The ensembles that are most widely studied in RMT, due to their applications in Physics and other fields, are the Gaussian ensembles. There are three basic classes of Gaussian ensembles [3]:

The Gaussian unitary ensemble $GUE(n)$ is described by the Gaussian measure with probability density,

$$\rho_{GUE} = \frac{1}{Z_{\mathrm{GUE(n)}}} exp\left(-\frac{n}{2}\mathrm{Tr}H^2\right) \tag{1.1}$$

defined on the space of $n \times n$ Hermitian matrices $H = (H_{\mathrm{ij}})_{\mathrm{i,j}=1}^{\mathrm{n}}$. Here $Z_{\mathrm{GUE(n)}} = 2^{n/2}\pi^{n^2/2}$ is a normalization constant, chosen so that the integral of the density is equal to one. The term unitary refers to the fact that the distribution is invariant under unitary transformation. The Gaussian unitary ensemble is used to model Hamiltonians lacking time-reversal symmetry.

The Gaussian orthogonal ensemble $GOE(n)$ is described by the Gaussian

measure with probability density

$$\rho_{GOE} = \frac{1}{Z_{\mathrm{GOE(n)}}} exp\left(-\frac{n}{4}\mathrm{Tr}H^2\right) \tag{1.2}$$

defined on the space of $n \times n$ real symmetric matrices $H = (H_{\mathrm{ij}})_{\mathrm{i,j}=1}^{\mathrm{n}}$. Its distribution is invariant under orthogonal transformation, and it is used to model Hamiltonians with time-reversal symmetry.

The Gaussian symplectic ensemble GSE(n) is described by the Gaussian measure with probability density

$$\rho_{GSE} = \frac{1}{Z_{\mathrm{GSE(n)}}} exp\left(-n\ \mathrm{Tr}H^2\right) \tag{1.3}$$

defined on the space of $n \times n$ quaternionic Hermitian matrices $H = (H_{\mathrm{ij}})_{\mathrm{i,j}=1}^{\mathrm{n}}$. Its distribution is invariant under transformation by the symplectic group, and it is used to model Hamiltonians with time-reversal symmetry, but no rotational symmetry.

One more important class of random matrices, often encountered in many real systems, are the Wishart matrices. These are $n \times n$ random matrices of the form

$$C = XX^* \tag{1.4}$$

where X is an $n \times n$ random matrix with independent random entries, and X^* is its conjugate matrix. In the important special case which was considered by Wishart, the entries of X are identically distributed Gaussian random variables, which are either real or complex. We will later use this class of matrices in a real system.

Empirical Correlation Matrices are important for the statistical analysis of real-world data. The empirical correlation matrices come from observations and measurements of real systems. They form a bridge between the real world and the mathematical formalism of RMT. Hence, study of statistical properties of such matrices is vital for the understanding of the underlying physical system. However, most real observations are contaminated by some kind of noise, which forms a significant component of the empirical correlation matrices. In order to understand the underlying physical system better, it is thus important to differentiate noise from actual useful information. It is a highly non-trivial problem to separate the useful components of the data from the noise. We will study more about this in Chapter 2.

A well-known technique used to differentiate the noise in the signal is to compare the statistical properties of the empirical correlation matrix with those of a correlation matrix obtained from uniformly distributed and independent random numbers. Many studies [2][4] have shown that most of

the eigenvalues of the empirical correlation matrices match with those of the random correlation matrices in RMT. This suggests that there is a considerable degree of randomness in the measured correlations. However, the few eigenvalues that deviate from the RMT values are the ones that contain some useful information about the system. The study of these deviating eigenvalues and their corresponding eigenvectors can provide valuable information about the system.

However, such an analysis is only possible when the data is non-categorical, i.e. in direct numerical form. For example, the values of heights of all students in a class, or the average annual rainfall measured at a particular location all constitute non-categorical data. In statistics, a categorical variable is a variable that takes a limited, and usually fixed, number of possible values. For example, the blood type of a person, the state or a geographical region in which a person resides, the political party he/she might vote for are all examples of categorical variables. For convenience in statistical processing, categorical variables can be assigned numeric indices, e.g. 1 through n for a n-way categorical variable (i.e. a variable that can be used to express exactly n possible values). In general, however, the numbers are arbitrary, and simply provide a convenient label for a particular value and have no significance. In other words, the values of a categorical variable exist on a nominal scale. They each represent a logically separate concept, cannot necessarily be meaningfully ordered, and cannot be otherwise manipulated as normal numbers could be.

As a result, it is not possible to analyse categorical data using standard methods in RMT. In order to do so, either the data has to be converted into non-categorical one, or some other method must be employed. Many a times, in real systems, we encounter categorical data. The main aim of the project is to devise a method that can tackle categorical data, to design a framework, analogous to RMT, that can be used as a background for analysis of real data.

In the next chapter, we discuss known basic results of RMT, including the method to calculate correlation matrices from data matrices. We present the theoretical expression for calculating the eigenvalue density of the correlation matrix in RMT. We also present the known theoretical expressions for the eigenvector distribution, information entropy and the nearest-level spacing distribution in RMT. These results form the backbone of statistical analysis using RMT.

In the third chapter, we discuss the problems encountered when dealing with categorical data. We then suggest two methods to compute random correlation matrices using categorical data. We study the statistical properties like eigenvalue density, spacing distribution and information entropy of

these random correlation matrices and contrast these results with the RMT predictions. These results now form the backbone for statistical analysis of categorical data.

In the fourth chapter, we demonstrate the usefulness of our method by applying it to two real-world systems:
1. Analysis of Indian General Elections results, where we analyse the data of previous Indian general elections and study the voting patterns for each of the 543 constituencies. We calculate correlations among various constituencies and study their statistical properties, in order to get some deeper insights into voting patterns across various constituencies.
2. Statistics of Atmospheric Pressure data, in which we analyse data containing atmospheric pressure values at different locations at different points of time. We find correlations among different locations and study the statistical properties of these correlation matrices. Since the data here is non-categorical, the analysis can also be done using RMT techniques. However, we apply our method here to show it can also be applied to non-categorical data.
Study of statistical properties of the correlation matrices like eigenvalue density, eigenvector distribution, information entropy and nearest level spacing distribution in each case gives valuable insight into the problems.

Chapter 2

Random Matrix Theory

2.1 Computing Correlations in RMT

In statistics, correlation constitutes of a broad class of statistical relationships involving dependence. Correlations are useful because they can possibly indicate a predictive relationship that can be exploited in practice. For example, study of correlations between prices of different stocks can be used to predict the stock prices at a certain time [2]. In multivariate statistics, correlations are usually depicted by correlation matrices, which for n random variables $X_1, X_2, ..., X_n$, is a $n \times n$ matrix whose ij^{th} entry is the correlation between X_i and X_j.

Correlation matrices are usually computed from data obtained from actual observations and measurements of various parameters. However, it is not always trivial to convert raw data, which is usually in the form of a time series, into a correlation matrix. In case the data consists of just one variable, such a calculation is trivial, as a univariate data only has one variance. However, in general, time series data is not always univariate. For instance, the measurement of atmospheric pressure at n different locations, with each of them varying with time, constitutes a multi-variate data.

The aim is to extract useful information from the multivariate data. A multivariate data with n variables will have $n(n+1)/2$ covariances. Thus, there is a lot more information and a lot more inter-dependencies among the n variables. However, this also means that there exist a possibility of a lot of redundancy in the data. Therefore, we need a transformation such that the transformed data has the desired properties.

Given a multivariate data x, we are looking for a transformation of the form [5],

$$y = Wx \tag{2.1}$$

where W is the transformation matrix with desired properties. We could require any of the two, depending on what we are looking for,(i) the transformation matrix will produce a resultant data y that captures the information in x or (ii) the transformed data y will capture the meaningful features in the data x. For the sake of our analysis, we will focus only on the first case, in which the transformation matrix has only the variance and the covariance information from x.

To fix ideas first, we will consider a case with bivariate data. We have time series of two variables, $u_i^{'}$ and $v_i^{'}$, $i = 1, 2, ..., p$. We assume that $p >> 2$ and both the data sets are centred as follows,

$$\eta = \eta^{'} - \langle \eta \rangle \tag{2.2}$$

where $\langle \eta \rangle$ is the sample mean of the data. Now, all possible covariances among (u, v) can be put in a matrix of the form,

$$C = \begin{pmatrix} c_{uu} & c_{uv} \\ c_{vu} & c_{vv} \end{pmatrix} \tag{2.3}$$

Here, c_{uv} is the covariance between the centered variables u and v and is given by,

$$c_{uv} = \sum_{i=1}^{p} u_i v_i \tag{2.4}$$

Note some of the important features of this matrix C. The diagonal elements are the variance of u and v. The off-diagonal elements are the covariances. We can similarly use correlation instead of covariances. In that case, we have the following property: Correlation of any variable with itself should be unity. Hence, the diagonal elements would be unity, while the off-diagonal elements will be $-1 \leq c \leq 1$. Of course, $c_{uv} = c_{vu}$ and thus C is a symmetric matrix.

The correlation matrix can be defined in terms of the data sets. Let Z be a matrix that is formed by assembling each time series as a column. For instance, in a bivariate case, we will have,

$$Z = \begin{pmatrix} u_1 & v_1 \\ u_2 & v_2 \\ : & : \\ u_p & v_p \end{pmatrix} \tag{2.5}$$

It is straightforward to see that,

$$C = \frac{1}{p} Z Z^T \tag{2.6}$$

This correlation matrix can be generalised to any number of variables.

2.2 Statistical properties of Correlation matrices

The correlation matrices calculated in the previous section, can in general be very complex. Extracting useful data from these matrices is highly non-trivial. Analysis of statistical properties of such matrices sheds light into useful information contained in these matrices.

2.2.1 Eigenvalue density

The density of eigenvalues in RMT is defined as [4]

$$\rho_C(\lambda) = \frac{1}{N}\frac{dn(\lambda)}{d\lambda} \tag{2.7}$$

where $n(\lambda)$ is the number of eigenvalues of C less than λ. If M is a $T \times N$ random matrix, $\rho_C(\lambda)$ is self-averaging and is exactly known in the limit $N \to \infty$ and $T \to \infty$ and $Q = \frac{T}{N} \geq 1$ fixed and reads [6]

$$\rho_C(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda}, \tag{2.8}$$

$$\lambda_{\min}^{\max} = \sigma^2(1 + 1/Q \pm 2\sqrt{1/Q}) \tag{2.9}$$

with $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, and where σ^2 is equal to the variance of the elements of M, equal to 1 in our normalization. In the limit $Q = 1$, the normalized eigenvalue density of the matrix M leads to the famous Wigner semi-circle law, and the corresponding distribution of the square of these eigenvalues, i.e., the eigenvalues of C is then given by Eq. (2.8) for $Q = 1$. The most important features predicted by Eq. (2.8) are as follows (see also Fig. 2.1):

(i) The lower edge of the spectrum is strictly positive (except for $Q = 1$); there are hence, no eigenvalues between 0 and $\lambda_{\min}$. Near this edge, the density of eigenvalues exhibits a sharp maximum, except in the limit $Q = 1(\lambda_{\min} = 0)$, where it diverges as $1/\sqrt{\lambda}$.

(ii) The density of eigenvalues also goes to zero above a certain upper edge $\lambda_{\max}$.

It should be noted that the above results are valid only in the limit $N \to \infty$. For finite N, the singularities present at both edges get smoothed, the edges become slightly blurred, with a small probability of finding eigenvalues above $\lambda_{\max}$ and below $\lambda_{\min}$, which goes to zero when N becomes large. The precise way in which these edges become sharp in the large N limit is actually known [7].

The eigenvalue density plot for $Q = 10$, $\lambda_{\min} = 0.47$ and $\lambda_{\max} = 1.73$ is shown in the figure 2.1.

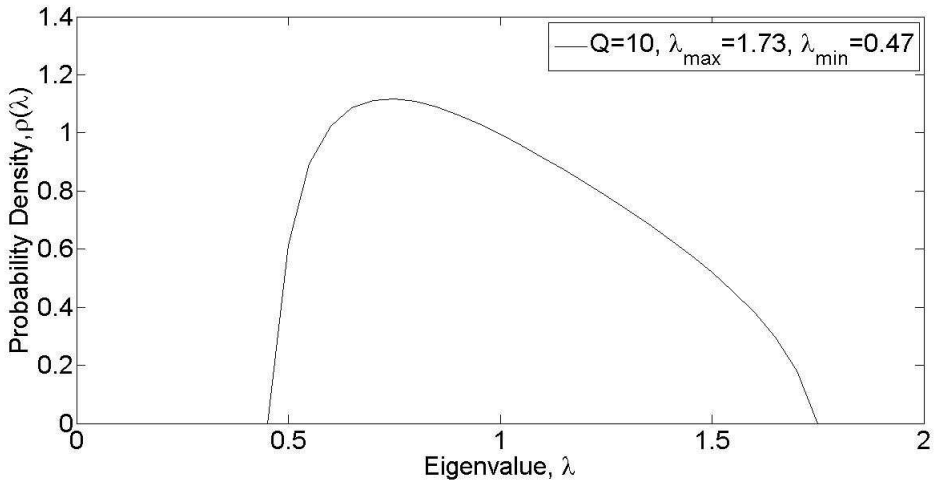

Figure 2.1: **The eigenvalue density plot predicted by RMT for Q=10**.

We can easily verify from the figure that there are no eigenvalues below $\lambda_{\min}$ and no eigenvalues above $\lambda_{\max}$. Depending on the system, we will have different eigenvalue density plots for various different values of Q, which in turn will depend on the size of the correlation matrix and the length of the time series.

2.2.2 Spacing distribution

One of the very famous results of the random matrix theory is the nearest-neighbour eigenvalue spacing distribution; i.e. the distribution of $s_{\mathrm{i}} = E_{\mathrm{i}+1} - E_{\mathrm{i}}$. It is the probability for finding the neighbouring levels with a given spacing s. The spacing distributions for the Gaussian Orthogonal Ensemble (GOE) and the Gaussian Unitary Ensemble (GUE) are given by [3],

$$P_{\mathrm{GOE}}(s) = \frac{\pi}{2} s \exp\left(-\frac{\pi}{4} s^2\right) \tag{2.10}$$

$$P_{\mathrm{GUE}}(s) = \frac{32}{\pi^2} s^2 \exp\left(-\frac{4}{\pi} s^2\right) \tag{2.11}$$

The distributions are plotted in Figure 2.2. For our analysis, we will use the GOE ensemble, unless specified otherwise.

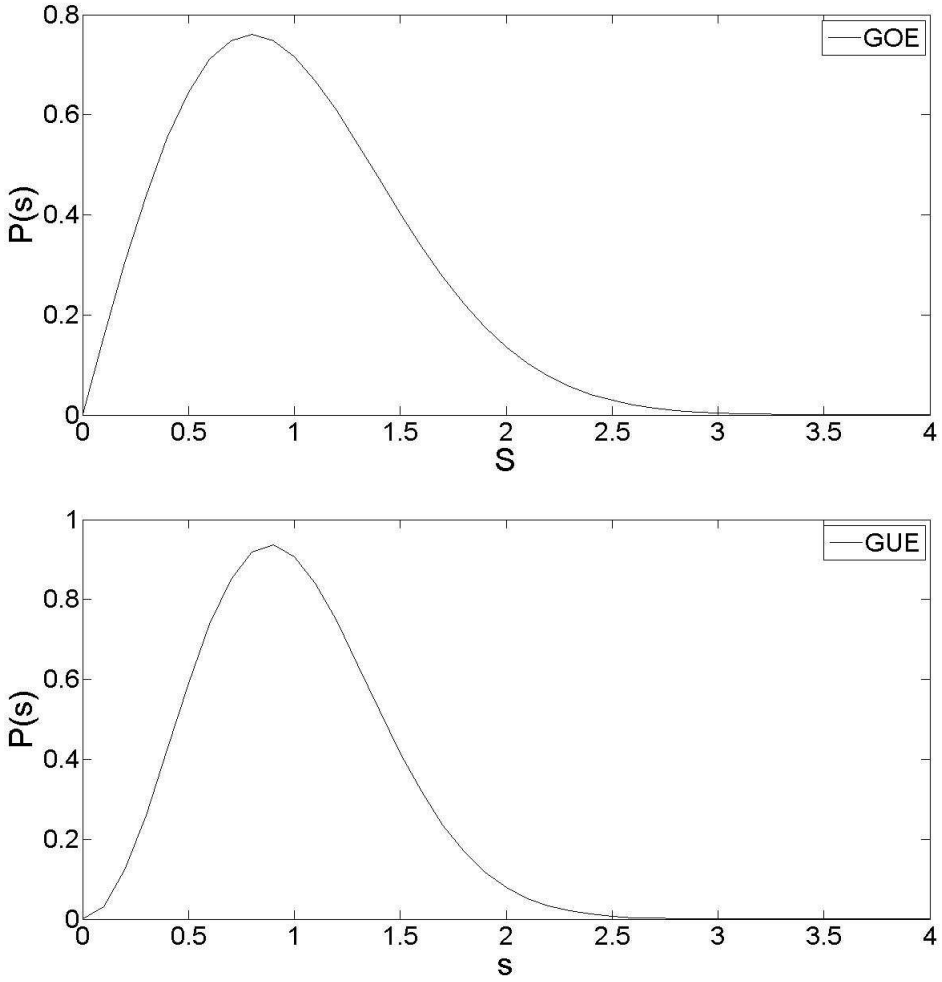

Figure 2.2: **Nearest neighbour spacing distribution for the GOE and GUE ensembles**.

The analytical forms above indicate level-repulsion, a tendency against clustering, as evident from low probability for small spacings. The level repulsion is linear for GOE and quadratic for GUE.

2.2.3 Eigenvector distribution

With the eigenvalue statistics alone, it is not straightforward to obtain detailed system specific information, unless there are significant deviations from random matrix predictions. The distribution of eigenvector components, on the other hand, reveals detailed and fine-grained information, at the level of every eigenvector.

Let a_j^m be the jth component of the mth eigenvector. Assuming that these components are Gaussian random variables with the norm being their only characteristic, it can be shown that the distribution of $r = |a_j^m|^2$, in the limit when the matrix dimension is large, is given by the special case of the χ^2 distribution [8]

$$P(r) = \left(\frac{\nu}{2\langle r\rangle}\right)^{\nu/2} \frac{r^{\nu/2-1}}{\Gamma(\nu/2)} exp\left(\frac{-r\nu}{2\langle r\rangle}\right) \tag{2.12}$$

The case $\nu = 1$ can be identified with GOE and gives the well-known Porter-Thomas (PT) distribution.

$$P(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \tag{2.13}$$

The distribution is plotted in Figure 2.3.

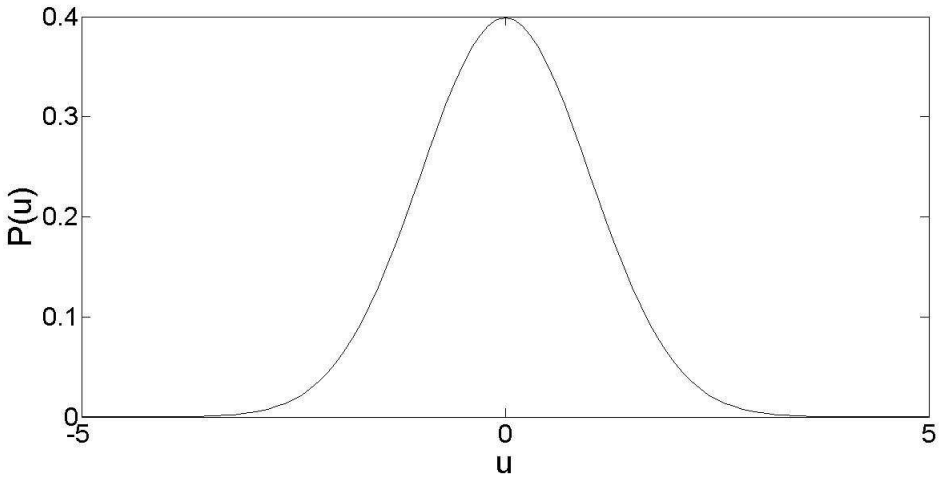

Figure 2.3: **The distribution of eigenvector components in GOE, also known as the Porter-Thomas distribution**.

The distribution of complex eigenvectors correspond to GUE class with $\nu = 2$. The general understanding is that if the eigenvectors are random, then its components are χ^2 distributed and deviations occur only if they show some symptoms of regularity.

2.2.4 Information Entropy

In information theory, entropy is a measure of the uncertainty in a random variable [9]. In this context, the term usually refers to the Shannon entropy,

which is is the average unpredictability in a random variable, which is equivalent to its information content. It also quantifies the expected value of the information contained in a message. For a discrete random variable X with possible values $x_1, ..., x_n$ and a probability mass function $P(X)$ taken from a finite sample, the entropy can be explicitly written as

$$H(X) = -\sum_i P(x_i) \log P(x_i) \tag{2.14}$$

In our analysis, $P(x_i)$ will be the square of the eigenvector components. For the Porter-Thomas distribution derived in the previous section, the Information Entropy can be expressed as [10]

$$H(1,1) = -\int_0^\infty y \log y P_\nu(y) dy = \log(d\nu/2) - \Psi(\nu/2+1) \tag{2.15}$$

To summarize, we have studied some basic results in Random Matrix Theory that are of great interest in understanding the statistical structure of the empirical correlation matrices. In the subsequent chapter, we demonstrate the problems encountered by RMT methods when the data is in non-numerical form, i.e. categorical data.

Chapter 3

Correlations of Categorical data

It often happens that the data obtained from a real system is not in direct numerical form, in a sense that we do not have numerical values as data points. Such a data is called categorical data. For example, consider a sample population of n individuals from p different geographical regions. You have to find correlations between different regions based on the blood groups of people from those regions. The empirical data that you have is the list of people, their region, and their corresponding blood groups, A, B, AB or O. There is no numerical value involved. Thus, it is not possible to use the RMT based method in this case for the statistical analysis.

The simplest method to convert this data into numerical values is to assign an index to each of the blood groups. For example, let A=1, B=2, AB=3 and O=4. Now, we have the data in numeric form, with a number corresponding to each individual. But, these numbers have no statistical significance as they are just indices and not actual values. The value of a number does not signify the weight of the quantity it denotes. For example, in normal numerical analysis, and the statistical analysis using RMT, the number 4 will have a higher statistical weight than the number 3. However, in our example, there is no reason why the blood group O should be favoured over the blood group AB. Thus, the RMT method for analysing categorical data might lead to a bias towards some quantity, which is not desirable. As we cannot use the RMT results for analysing correlations of categorical data, we have to devise a new method for it. The new method must be able to provide a theoretical framework to compute the correlations and analyse its statistical properties, analogous to the one provided by RMT. The statistical properties of the empirical correlation matrices computed for categorical data should then be compared with those of the theoretical one, just as we compare properties of non-categorical empirical correlation matrices with their corresponding RMT predictions.

3.1 Our approach

We present here two ways to compute correlations of categorical data. Both ways yield correlation matrices with identical statistical properties and hence, any one of them can be used for statistical analysis, depending on the system. Using these methods, we calculate random correlation matrices. We then study the statistical properties of these matrices like the eigenvalue density, eigenvector distribution and the spacing distribution. These results will then be used as a background for analysis of categorical data.

3.1.1 Uniformly distributed random numbers

p objects (denoted by numbers $1-p$) are uniformly distributed in a matrix A of size (n,t), where $t = 1,2,3,.....,T$ and $n = 1,2,3,.....,N$. Here, t is analogous to a time series in real data, while n is analogous to the number of nodes in a network which represents the data. In real data, n can be anything, like spatial locations, financial stocks of various companies, etc.
The algorithm for calculating correlations is as follows:
Every row of the matrix A is compared with every other row, term by term and the results are stored in an intermediate matrix B of size (n^2,t). The entries of B can be obtained as follows:
Write

$$i = nq + r \qquad (0 \leq r < n) \tag{3.1}$$

Now, if $r = 0$, define

$$\begin{aligned} B_{ij} &= 0 \qquad \text{if } A_{qj} \neq A_{nj} \\ &= 1 \qquad \text{if } A_{qj} = A_{nj} \end{aligned} \tag{3.2}$$

If $r \neq 0$, define

$$\begin{aligned} B_{ij} &= 0 \qquad \text{if } A_{(q+1)j} \neq A_{rj} \\ &= 1 \qquad \text{if } A_{(q+1)j} = A_{rj} \end{aligned} \tag{3.3}$$

Every row of B is then averaged and after reshaping, we get a correlation matrix C of size (n,n). This correlation matrix is, by construction, symmetric and all its diagonal entries are 1.

3.1.2 Dividing numbers into intervals

We generate a list of n uniformly distributed random numbers in the interval $(0,1)$. We then form t bins of equal size and allot each of the numbers in the

above list to a bin. Thus, we now have a series of numbers (1 to p) distributed in a $n \times t$ matrix A. The correlation matrix C is then computed using the technique used in the previous section, viz. row by row comparison. This correlation matrix too is, by construction, symmetric and all its diagonal entries are 1.

Mathematically, if $x(i)$ and $y(i)$ are two row vectors (in this case, two time series), each containing n elements, then the correlation C_{xy} between them is given by:

$$C_{\mathrm{xy}} = \frac{\sum_{i=1}^{n} \delta_{x_i,y_i}}{n} \tag{3.4}$$

In some cases, we also take weighted average to compute the correlation matrix, with a weight w_i associated with the ith element of x and y. In that case, C is given by:

$$C_{\mathrm{xy}} = \frac{\sum_{i=1}^{n} w_i \delta_{x_i,y_i}}{\sum_{i=1}^{n} w_i} \tag{3.5}$$

To see how these correlation matrices are computed, consider the following simple example: Let

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Here, there are just $p = 2$ objects (1 and 0), the length of the time series is $T = 5$ and the number of nodes is $N = 2$. Now let us compute the correlation by row by row comparison and construct the intermediate matrix B. To compute the first row of B, let us compare the first row of A with itself. Here, $i = 1$, $n = 2$. Thus, $q = 0$ and $r = 1$. Hence, from Eq. 3.3, we can see that, $\forall j$,

$$\begin{aligned} B_{1j} &= 0 \qquad \text{if } A_{1j} \neq A_{1j} \\ &= 1 \qquad \text{if } A_{1j} = A_{1j} \end{aligned}$$

Obviously, $A_{1j} = A_{1j}\ \forall j$, and hence the first row of B will be $\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}$. The same is true when we compare the second row of A with itself. Thereby, the fourth row of B is also $\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}$.

To get the second (and hence, third) row of B, we compare the first row of A with the second. Here, $i = 2$ and $n = 2$. Thus, $q = 1$ and $r = 0$. Hence,

from Eq. 3.2, we can see that, $\forall j$

$$\begin{aligned} B_{2j} &= 0 \qquad \text{if } A_{1j} \neq A_{2j} \\ &= 1 \qquad \text{if } A_{1j} = A_{2j} \end{aligned}$$

Varying j from 1 to 5, we observe that the second, third and fourth entries match, while the first and fifth don't. Hence, $B_{21} = 0$, $B_{22} = 1$, $B_{23} = 1$, $B_{24} = 1$ and $B_{25} = 0$. Thus, the second and third rows of B are equal to $\begin{pmatrix} 0 & 1 & 1 & 1 & 0 \end{pmatrix}$. We thus have,

$$B = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Averaging every row of B, we get a 4×1 matrix,

$$D = \begin{pmatrix} 1 \\ 0.6 \\ 0.6 \\ 1 \end{pmatrix}$$

The entries of D can be interpreted as correlations between the different rows of A. On reshaping D, we get a symmetric 2×2 correlation matrix C.

$$C = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$$

The entries in C, i.e., the correlations, take values in the interval $(0, 1)$. Here, a correlation of 1 indicates that the two quantities are completely correlated, while a correlation of 0 indicates that the they are completely anti-correlated. Unlike the Pearson Correlation Coefficient, which takes values from -1 to 1, there are no negative correlations here. The diagonal elements are the correlations of each quantity with itself, which obviously should be equal to 1, as is the case.

The random correlation matrices computed by both the methods are statistically identical, as we will see in the next section, when we study various statistical properties of C. For our analysis, we will construct a random matrix with $n = 543$ and $t = 32$. The reason for choosing these particular values comes from a real application, analysis of Indian Elections data, which we will study in detail in Chapter 4. Depending on the data matrix of a particular system, a similar analysis can be done for that corresponding size of the matrix. For now, we will just list the statistical result for the correlation matrix of size $(543, 543)$.

3.2 Statistical Properties of C

3.2.1 Number of non-zero eigenvalues

The number of non-zero eigenvalues is equal to the rank of the correlation matrix. The rank of a matrix is equal to the number of linearly independent rows (or columns) of the matrix, which will depend on various factors such as the number of nodes and number of independent objects (entries). For the case of the correlation matrix computed using any of the two methods above, the following are observed:
1. More the number of possible entries, more is the number of linearly independent rows and columns, and hence, more is the rank. Thus, the number of non-zero eigenvalues increases with the number of possible entries, i.e. number of objects p, while keeping the number of nodes and the length of the time series fixed.
2. The number of non-zero eigenvalues decreases with the increase in the number of nodes n, which is equal to the number of rows of the matrix, while keeping the number of objects constant and the length of the time series constant. This is expected as increasing the number of rows makes it less likely to increase the number of linearly independent rows, thereby decreasing its rank.
3. The number of non-zero eigenvalues increases with the increase in the length of the time series t, while keeping the number of objects and the number of nodes fixed.

3.2.2 Variation of $\lambda_{\max}$

The maximum eigenvalue $\lambda_{\max}$ of the random correlation matrix is statistically significant as it might be the one that contains the most useful information. Hence, the study of $\lambda_{\max}$ and its statistical properties is very important for our analysis. We study the variation of $\lambda_{\max}$ with respect to the number of objects, the number of nodes and the length of the time series.

$\lambda_{\max}$ vs No. of objects

The maximum eigenvalue $\lambda_{\max}$ of the random correlation matrix exponentially decreases with increase in the number of objects, while keeping the number of nodes and the length of the time series constant. The plot of $\lambda_{\max}$ vs no. of objects is shown in Figure 3.1.

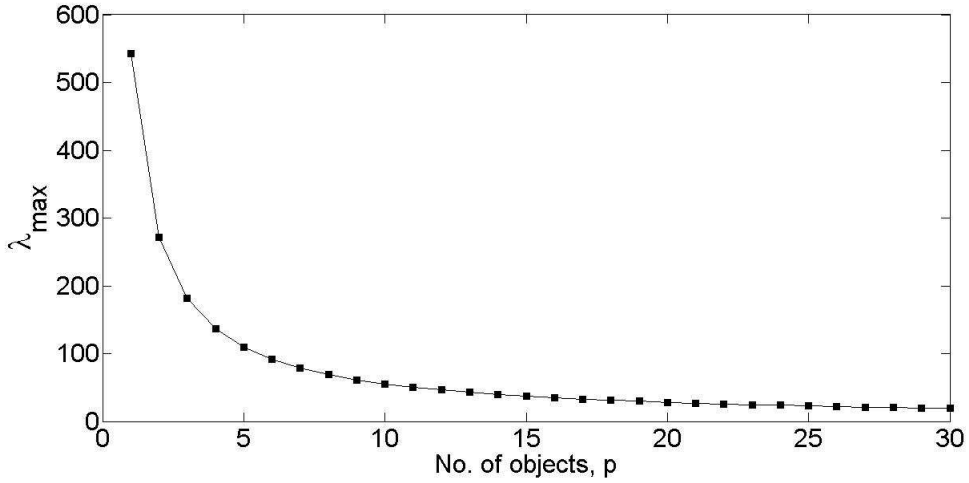

Figure 3.1: **Variation of λ_{max} with no. of objects**. We can see that λ_{max} decays exponentially with increase in the number of objects.

λ_{max} vs Length of the time series

The maximum eigenvalue λ_{max} remains constant with the length of the time series, while keeping the number of nodes and the number of objects constant. The plot of λ_{max} vs t is shown in Figure 3.2.

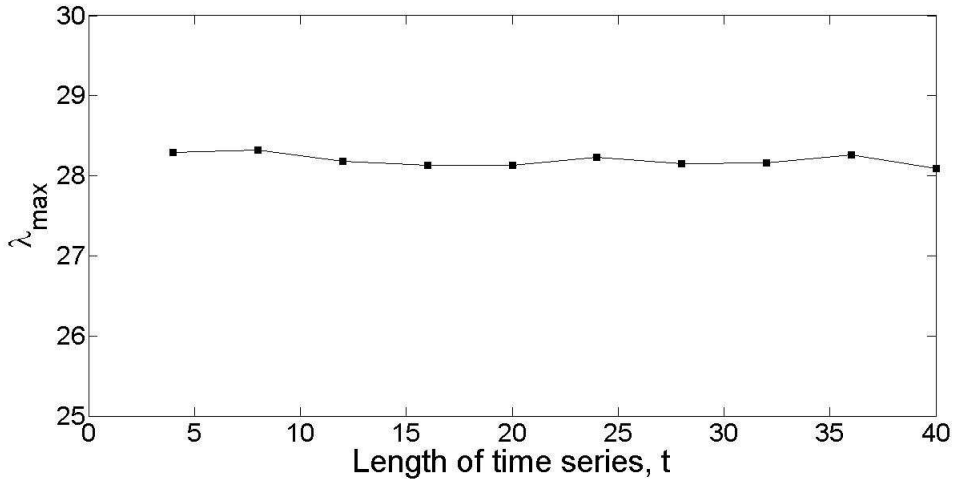

Figure 3.2: **Variation of λ_{max} with the length of the time series**

As we can see from Figure 3.2, there is a variation of less than 0.25 in the value of λ_{max}, even as t varies from 0 to 40.

λ_{max} vs number of nodes

The maximum eigenvalue λ_{max} increases linearly with the number of nodes n, while keeping the number of objects and the length of the time series fixed. The plot of λ_{max} vs n is shown in Figure 3.3.

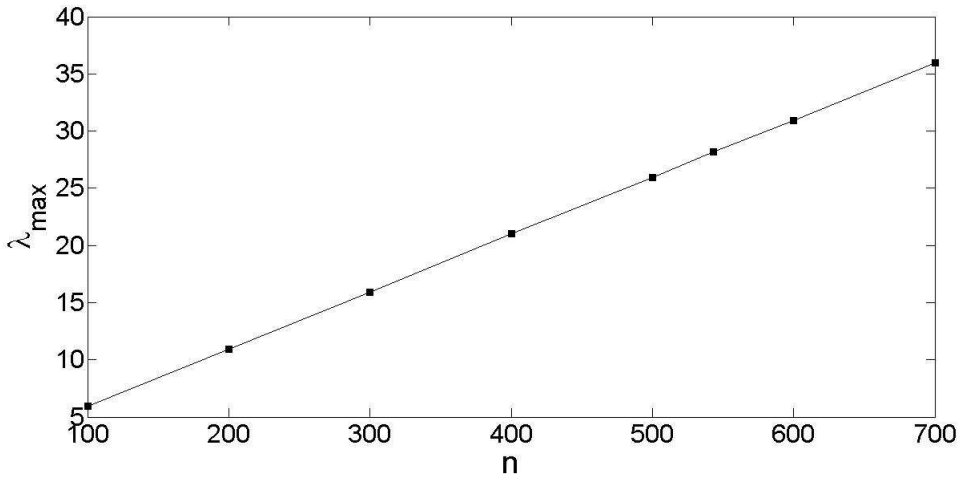

Figure 3.3: **Variation of λ_{max} with the number of nodes**. A linear plot is observed.

As it is evident from Figure 3.3, the plot is almost a straight line passing through the origin, with increments of 5 units in the value of λ_{max} with every 100 units increment in the value of n.

3.2.3 Eigenvalue density

The distribution of eigenvalues of the correlation matrix can provide valuable information about the system. Comparing the eigenvalue density plot of the empirical correlation matrix with that of the random correlation matrix can help distinguish between actual data and noise in the system. The eigenvalue density plot for the correlation matrix computed by our approach, for both the methods, viz. uniformly distributed random numbers and random numbers divided into intervals, is shown in figure 3.4. As can be seen, both the methods yield identical plots for the eigenvalue density of C. Henceforth, depending on the system, we will use any one of them, whichever is suitable, for our analysis.

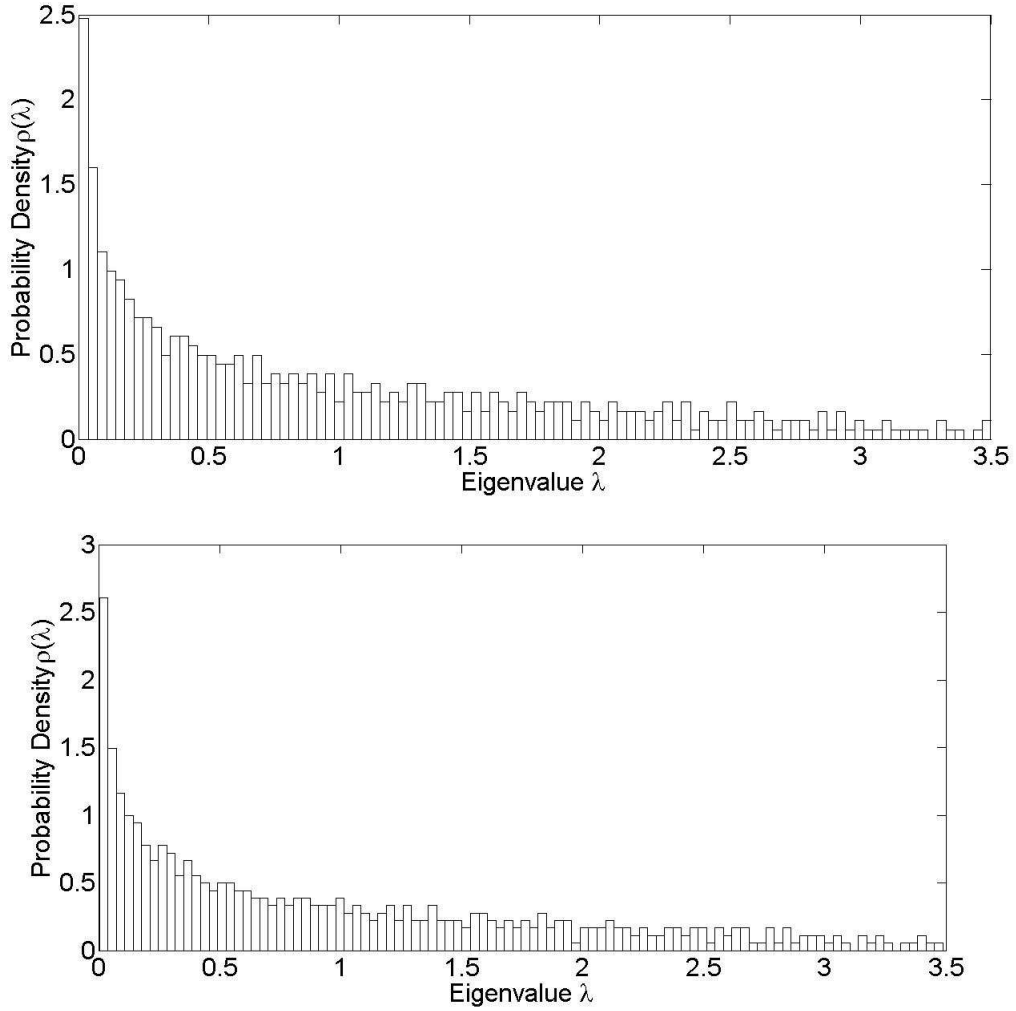

Figure 3.4: **Eigenvalue density plots of C for a)Uniformly distributed random numbers b)Dividing random numbers into intervals.**.

Note that these plots are only for a particular size of the correlation matrix C, (543×543). The plot will change as we vary the size, depending on the data. In spite of that, the eigenvalue density plot for our approach differs quite a bit from the one predicted by RMT. This is not unexpected, as our method handles categorical data, while the RMT one deals only with non-categorical one. However, there is no known analytical expression for the eigenvalue density, analogous to Eq. 2.8 in RMT. The analysis for finding such an expression is in progress.

3.2.4 Spacing Distribution

We next consider the eigenvalue spacing distribution, which reflects two point as well as eigenvalue correlation functions of all orders. We compare the eigenvalue spacing distribution of the random matrix C with that of GOE random matrices, given by Eq. 2.10, which is also referred to as the Wigner Sunrise [3]. The nearest neighbour spacing distribution for both of our methods is shown in the figure 3.5:

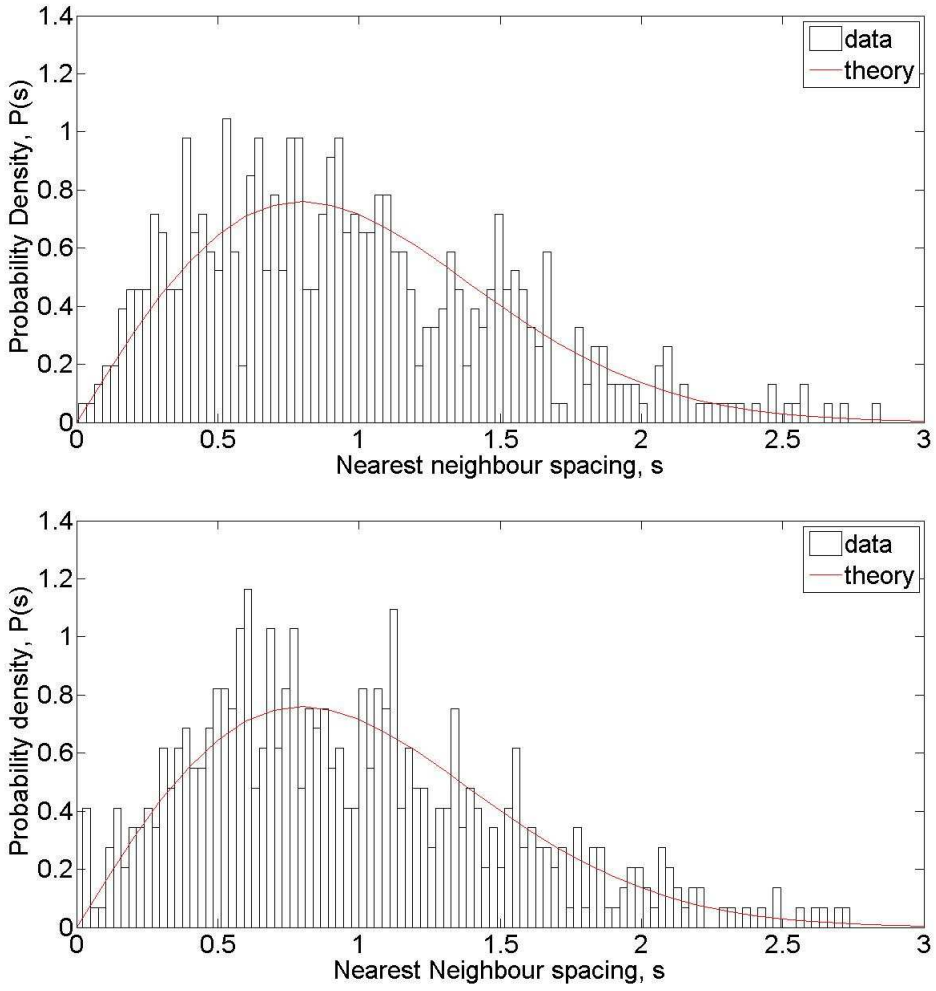

Figure 3.5: **Nearest neighbour spacing for a)Uniformly distributed random numbers b)Dividing random numbers into intervals. The solid curve shows the GOE prediction**.

Again, we see that both the approaches yield identical spacing distribution plots. Additionally, we also see a nice agreement between our approach

and the GOE predictions in RMT.

3.2.5 Eigenvector Distribution

The eigenvector distribution for the correlation matrix is shown in the figure 3.6.

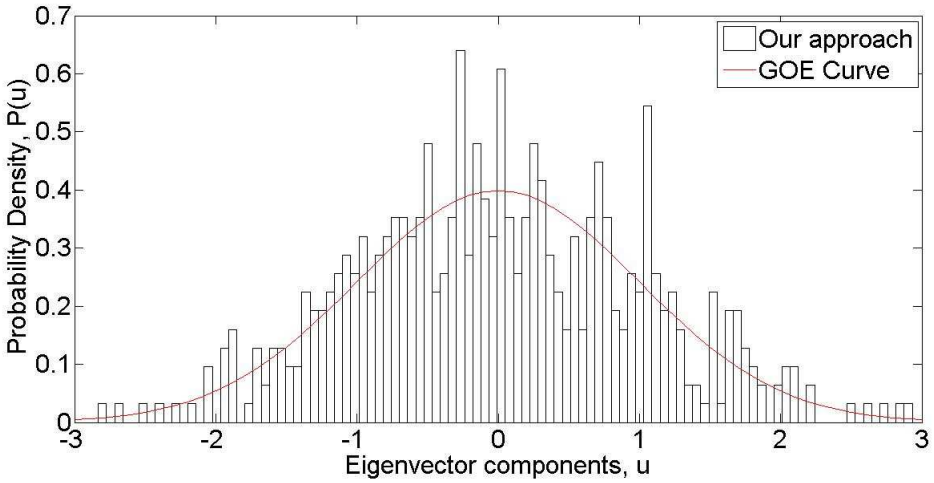

Figure 3.6: **Eigenvector distribution for C. The solid curve shows the GOE prediction**.

As we can see, the histogram of the eigenvector components of the correlation matrix C is a Gaussian and fits reasonably well with the GOE prediction. This shows that our approach can be used for the study of the statistical properties of empirical correlation matrices constructed from categorical data matrices.

3.2.6 Information Entropy

The information entropy H of the system is computed using the formula [9]:

$$H_i = -\sum_j a_{ij} \log a_{ij} \tag{3.6}$$

where a_{ij} are actually the square of the eigenvector components. The plot is shown in the Figure 3.7.

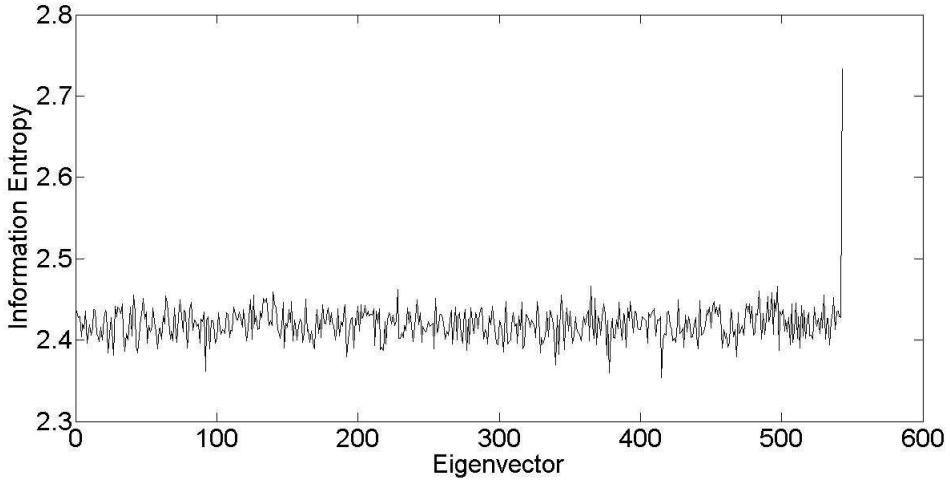

Figure 3.7: **Information Entropy of the square of eigenvector components in our approach.**

We can see that the information entropy is more or less constant for all the components, except the last one. This shows that the data predominantly consists of random correlations.

In this chapter, we have developed a new approach to compute correlations and analysed and compared its statistical properties with those in RMT. This results can now be used as a background for analysing real-world data, which will be done in the next chapter.

Chapter 4

Applications to real systems

The motivation for the correlation calculation for categorical data comes from the analysis of real data, as observations of many real systems does not always yield non-categorical data. Empirical correlation matrices are of great importance in data analysis in order to extract the underlying information contained in experimental signals and time series. Real-life data is often contaminated with noise. It is thereby vital to be able to separate the noise from the signal in order to extract some useful information from the data. It is thus imperative to devise methods which helps one to distinguish signal from noise. In other words, there should be a method to distinguish eigenvectors and eigenvalues of the correlation matrix containing real information from those which are devoid of any useful information. From this point of view, it is interesting to compare the statistical properties of an empirical correlation matrix C to a purely random matrix, sort of a null hypothesis, obtained from a finite time series of independent and identically distributed random numbers. Deviations from the random matrix case might then give clues about the presence of significant information.

In this section, we use the random correlation matrices computed by our method in the previous chapter as a null hypothesis and compare their statistical properties to those of empirical correlation matrices obtained from the following two real systems:
1. Indian Elections Data
2. Atmospheric Pressure data
We try to see if any useful information can be extracted from this analysis, which will give better insights into these systems.

4.1 Indian Elections Data

With its multi-party democracy, the Indian political system is amazingly complex. The presence of many regional parties, in addition to various diverse factors such as caste, religion, region and language makes the analysis of poll outcomes increasingly difficult. With the biggest ever and one of the most anticipated general elections scheduled in April-May 2014, it is indeed exciting to analyse how different constituencies are correlated to each other, in a sense that whether results or trends in one constituency affect those in the other constituencies. Few studies have already tried to analyse the outcome of elections in various countries [11][12]. Our original intention was to come up with a prediction of the outcome of the general elections, based purely on previous results, unlike most poll predictors that involve opinion polls. However, due to the incredible complexity involved due to a large number number of factors affecting the poll outcome, it is nearly impossible to predict the results with certainty. Nonetheless, it is still possible to extract some useful information, in terms of finding the correlations between the voting patterns of different constituencies in the country. In this section, we show that our approach to compute correlations for categorical data can be used for the analysis of the Indian elections.

For our analysis, we choose the data of the previous seven Indian General Election results (from 1984 to 2004), which was obtained from the Election Commission of India website (http://eci.nic.in/). The reason to start from 1984 elections and not from the first general election in 1952 is that prior to 1984, the Indian political space was mostly dominated by the Congress. There was no significant opposition and the number of regional parties was also less. This leads to bias towards the Congress party in the analysis, which is not desirable in today's political scenario, where the regional parties enjoy a significant clout in their respective states.

After the data collection, we identify 18 major parties and assign a number to each party. All the other minor parties are clubbed as one unit, and given a number 19. For example, Congress is assigned the number 1, while BJP is assigned 2. Small parties like PDMK, RLD, etc. are all clubbed together in 19. We then obtain a time series (of $T = 7$ steps, signifying the 7 elections) for each of the 543 constituencies, giving us a 543×7 data matrix A. Every time series, i.e. row of A, is then compared with every other time series. If in any particular year, two constituencies i and j have elected the same party, we assign $A(ij) = 1$. If they have elected different parties, we assign $A(ij) = 0$. This gives us an intermediate matrix B as per Eq. 3.2 and Eq. 3.3, consisting of 1s and 0s. Every row of B is then averaged using weighted average as per Eq. 3.5, with more weight to the more recent

election, to calculate the correlation between each constituency.

The reason for using weighted average instead of normal average is that the more recent elections will have more influence and relevance to present voting patterns than the older elections. This gives us a correlation matrix C of size (543,543). This was precisely the reason why the size of the random correlation matrix in Chapter 3 was $n = 543$. The correlation matrix sheds light on various hidden correlations between various constituencies in different regions. More useful information can be extracted by computing the eigenvalues and eigenvectors of the correlation matrix. The normalised eigenvalue density plot is shown in Figure 4.1.

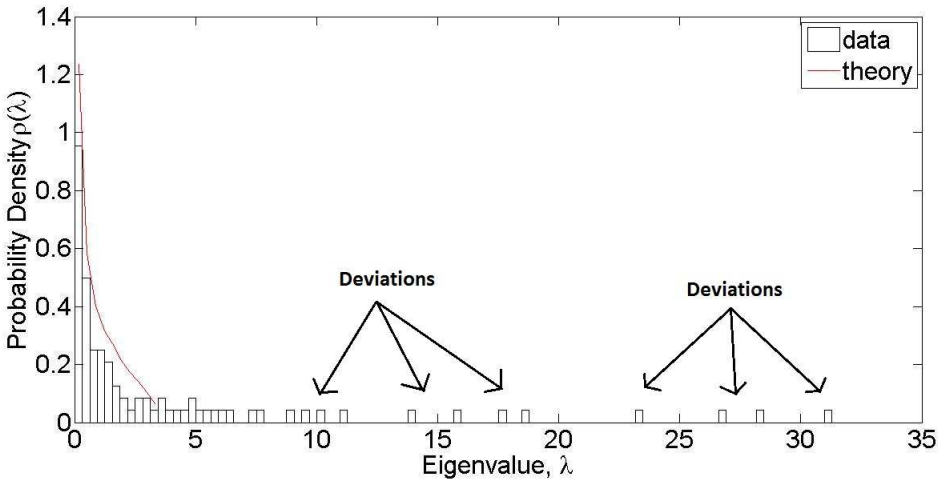

Figure 4.1: **Eigenvalue Density plot for Indian elections data. The solid curve is the density for the corresponding random matrix.**

It can be observed that while many of the eigenvalues nearly coincide with those of the corresponding random matrix, few of them deviate. Those which coincide can be accounted for by the noise in the data, while those eigenvalues, and their corresponding eigenvectors that deviate are the ones that actually contain useful information about the system. Thus, it is imperative to focus attention on those eigenvectors and perform advanced statistical analysis in order to extract the information. However, such an analysis is by no means trivial, requiring some advanced statistical tools and hence, we will not cover that here.

The nearest neighbour spacing distribution is shown in the Figure 4.2.

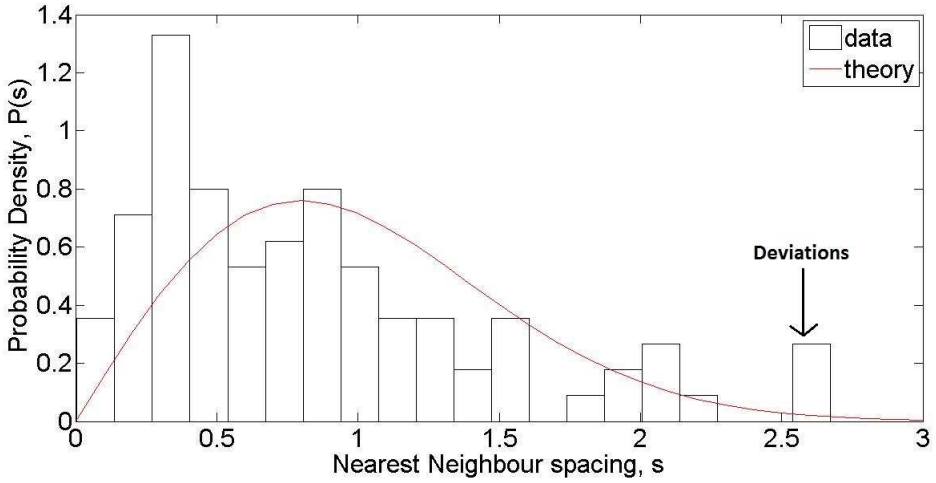

Figure 4.2: **Nearest neighbour spacing distribution for the Indian general elections data. The solid curve shows the GOE prediction**.

We see a good fit with the GOE curve, with a few deviations, thereby confirming the fact that most of the data is dominated by random correlations. The deviations suggest the presence of significant information.

The distribution of eigenvector components is shown in the Figure 4.3.

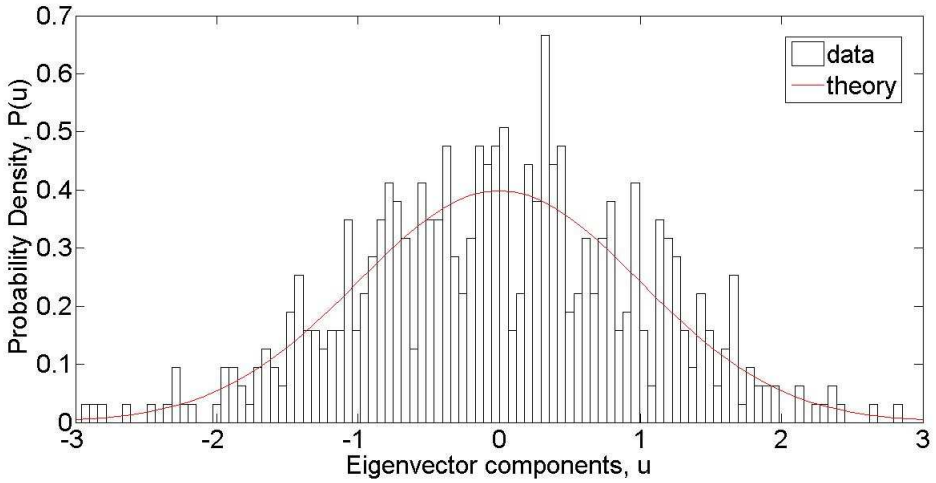

Figure 4.3: **Distribution of eigenvector components for data from Indian Elections. The solid curve is the GOE prediction.**

As can be observed from the figure, the distribution of eigenvector components agrees quite well with the GOE curve predicted by RMT.

4.2 Atmospheric Pressure Data

The state of the Earth's atmosphere is governed by the classical laws of fluid motion. There exist a great deal of correlations in various spatial and temporal scales in the atmosphere. These correlations are critical to understand the short and long term trends in climate. Generally, it is possible to recognise atmospheric correlations from the study of empirical correlation matrices constructed from the atmospheric data of parameters such as temperature, pressure, etc. Most significant correlations are documented as teleconnection patterns, which are the simultaneous correlations in the fluctuations of large scale atmospheric parameters, measured at widely separated points on the earth.

Empirical correlation matrices are widely used in atmospheric sciences, for example, to analyse the large scale patterns of atmospheric variability. Various computational methods, based on the Monte-Carlo simulations [13] have been used for the purpose of separating noise from the data and extract actual physical information from it. Beyond a point, these methods become computationally expensive and are replaced by asymptotic formulations [13]. Atmospheric correlations arise naturally from known physical interactions and hence are interesting to study from a RMT perspective because they offer instances to verify two (orthogonal and unitary) of the three Gaussian ensembles of RMT discussed in Chapter 2. The random matrix analysis can be successfully applied to empirical correlation matrices obtained from the analysis of the basic atmospheric parameters that characterise the state of the atmosphere. These methods turn out to be very useful as a tool to separate the signal from the noise, with lesser computational expense than with methods based on Monte-Carlo techniques.

Santhanam and Patra [14] have shown that significant information can be obtained from the study of empirical correlation matrices. In general, any atmospheric parameter $z(x,t)$, (like wind velocity, geopotential height, temperature etc.),varies with space(x) and time(t) and is assumed to follow an average trend on which the variations are superimposed, i.e.,

$$z(x,t) = z_{\text{avg}}(x) + z'(x,t) \tag{4.1}$$

If the observations were taken t times at each of the n spatial locations and the information is contained in the data matrix Z of order n by t, then the correlation matrix is given by

$$C = \frac{1}{n} Z Z^T. \tag{4.2}$$

Analysis of statistical properties of C, like the nearest neighbour spacing

distribution and eigenvector components, and comparing them with RMT predictions, gives significant physical insight into the problem.

Although the above analysis is done using RMT technique, we demonstrate that the same analysis can be done using our method as well, thereby showing that our method is applicable for non-categorical data as well. The data used is the same as the one used in [14]. We have a data matrix with $n = 434$ and $t = 624$. The raw data is first normalised by subtracting the mean and dividing by the standard deviation. The normalised data is then divided into 20 intervals of equal sizes. This gives us a new matrix B, which consists of numbers from 1-20, for each spatial location. The correlation matrix C is then computed using the familiar row-by-row comparison technique developed in Chapter 3.

As done for the elections data, we now study some statistical properties of the empirical correlation matrix computed above. The eigenvalue density plot is shown in Figure 4.4.

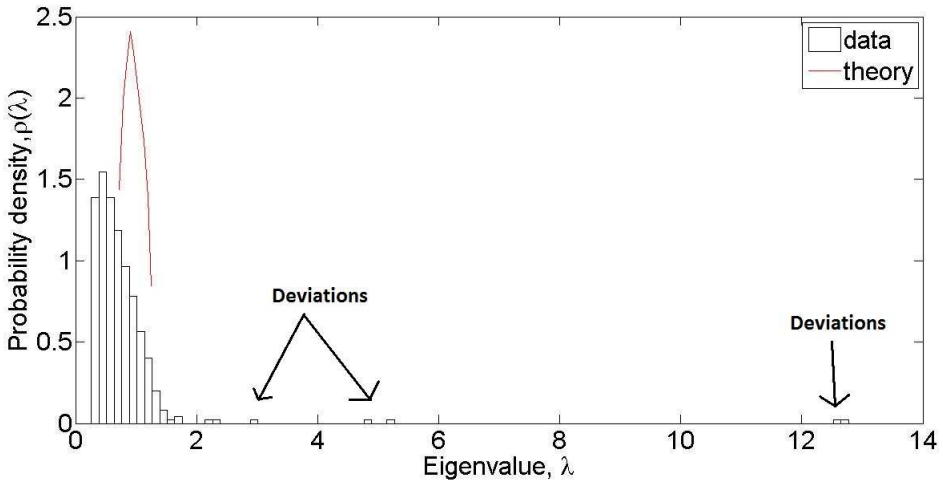

Figure 4.4: **The eigenvalue density plot for the atmospheric pressure data. The solid curve shows the density of the corresponding random correlation matrix.**

It can be seen that the bulk of the eigenvalues fall within the predicted values of the corresponding random matrix, indicating the domination of random correlations in the data. Study of the ones that deviate, along with their corresponding eigenvectors, might indicate the presence of significant information about the system.

The nearest neighbour spacing distribution is shown in Figure 4.5.

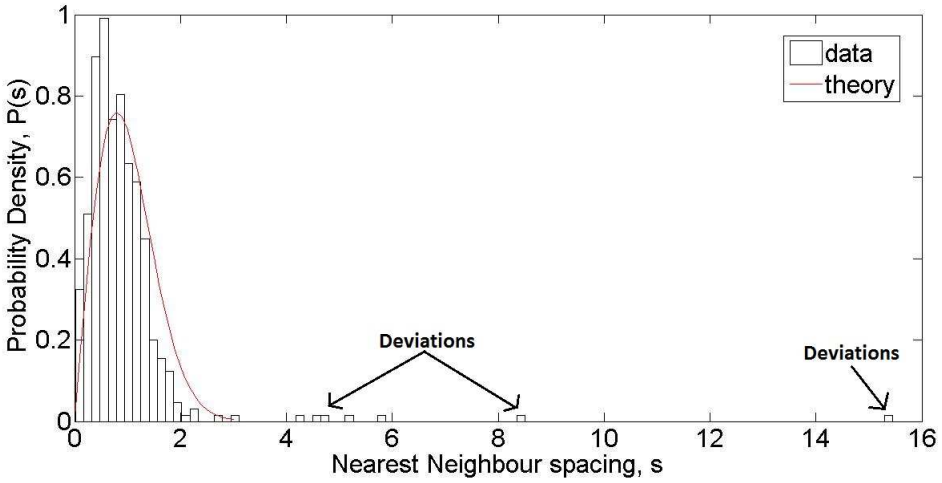

Figure 4.5: **Nearest neighbour spacing distribution for the atmospheric pressure data. The solid curve shows the GOE prediction.**

As we can see from the figure, the curve obtained using our method fits perfectly with the GOE curve, except for a few deviations, which indicate the presence of true information.

The eigenvector distribution is depicted in Figure 4.6.

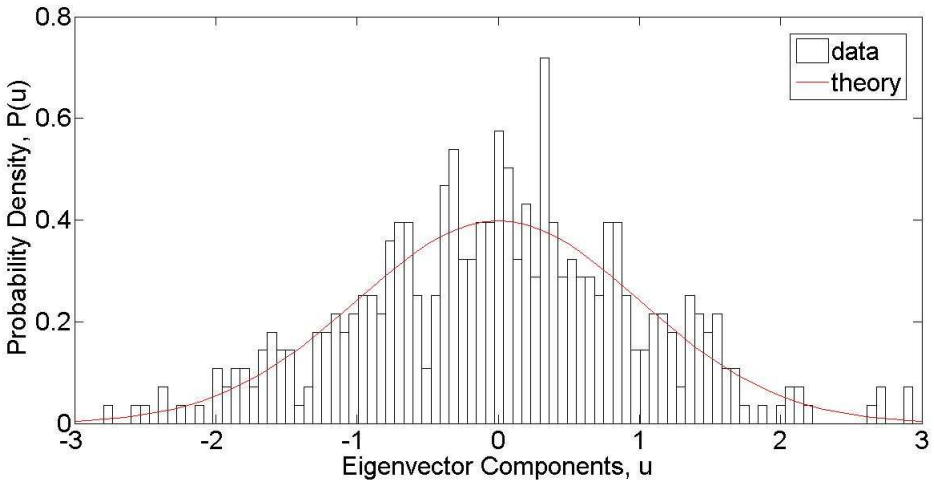

Figure 4.6: **Distribution of eigenvector components. The solid curve is the GOE prediction.**

Again, we see a reasonable agreement with the GOE predictions.

To summarize, in this chapter, we applied our method to compute correlation to two different systems and studied the statistical properties of the

empirical correlation matrices of these systems. In both the cases, we found that the statistical properties of eigenvalue density, eigenvector distribution and the nearest neighbour spacing distribution match quite well with the theoretical values, thus signifying that the method is indeed useful in the analysis of categorical data.

Chapter 5

Conclusion and Further Scope

This work shows that our approach to compute correlation matrices for categorical data can be used for statistical analysis of empirical correlation matrices arising in many real-world systems, which is not possible using the known methods in Random Matrix Theory. This work lays down a background theoretical framework against which the statistical properties of empirical correlation matrices can be compared. It happens that most statistical properties of the empirical correlation matrices, like the eigenvectors and eigenvalue, match with those of the corresponding random ones, except a few deviating ones, thereby denoting the presence of noise in the data. It is the study of these eigenvectors that can provide valuable information about the system.

Further scope for this work is to employ advanced statistical methods for the analysis of such deviating eigenvectors, as has been studied in the context of financial data in [4]. One more challenging problem is to find an analytical form for the eigenvalue density for our approach, analogous to the one in RMT. We are looking to carry this work forward and hope to address these problems soon.

References

[1] E. P. Wigner, Characteristic vectors of bordered matrices with infinite dimensions i, in: The Collected Works of Eugene Paul Wigner, Springer, 1993, pp. 524–540.

[2] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, H. E. Stanley, Random matrix approach to cross correlations in financial data, Physical Review E 65 (6) (2002) 066126.

[3] M. L. Mehta, Random matrices, Vol. 142, Academic press, 2004.

[4] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, Noise dressing of financial correlation matrices, Physical review letters 83 (7) (1999) 1467.

[5] M. Santhanam, Multivariate statistics (2004).
URL `http://www.iiserpune.ac.in/~santh/mvstat.pdf`

[6] A. Sengupta, P. P. Mitra, Distributions of singular values for some random matrices, Physical Review E 60 (3) (1999) 3389.

[7] M. J. Bowick, É. Brézin, Universal scaling of the tail of the density of eigenvalues in random matrix models, Physics Letters B 268 (1) (1991) 21–28.

[8] F. Haake, Quantum signatures of chaos, Vol. 54, Springer, 2010.

[9] S. Ihara, Information theory for continuous systems, Vol. 2, World Scientific, 1993.

[10] K. Jones, Entropy of random quantum states, Journal of Physics A: Mathematical and General 23 (23) (1990) L1247.

[11] M. Hanias, L. Magafas, Application of physics model in prediction of the hellas euro election results., Journal of Engineering Science & Technology Review 2 (1).

[12] H. Hernández-Saldaña, Three predictions on july 2012 federal elections in mexico based on past regularities, arXiv preprint arXiv:1207.0078.

[13] R. W. Preisendorfer, F. Zwiers, T. Barnett, Foundations of principal component selection rules, SIO Reference Series 81-4 May 1981. 192 p, 37 Fig, 33 Tab, 75 Ref.

[14] M. Santhanam, P. K. Patra, Statistics of atmospheric correlations, Physical Review E 64 (1) (2001) 016102.