4

Wishart Ensemble and Marčenko–Pastur Distribution

In this chapter we will study the statistical properties of large sample covariance matrices of some N-dimensional variables observed T times. More precisely, the empirical set consists of $N \times T$ data $\{x_i^t\}_{1 \le i \le N, 1 \le t \le T}$, where we have T observations and each observation contains N variables. Examples abound: we could consider the daily returns of N stocks, over a certain time period, or the number of spikes fired by N neurons during T consecutive time intervals of length Δt, etc. Throughout this book, we will use the notation q for the ratio N/T. When the number of observations is much larger than the number of variables, one has $q \ll 1$. If the number of observations is smaller than the number of variables (a case that can easily happen in practice), then $q > 1$.

In the case where $q \to 0$, one can faithfully reconstruct the "true" (or population) covariance matrix $\mathbf{C}$ of the N variables from empirical data. For $q = \mathcal{O}(1)$, on the other hand, the empirical (or sample) covariance matrix is a strongly distorted version of $\mathbf{C}$, even in the limit of a large number of observations. This is not surprising since we are trying to estimate $\mathcal{O}(N^2/2)$ matrix elements from $O(NT)$ observations. In this chapter, we will derive the well-known Marčenko–Pastur law for the eigenvalues of the sample covariance matrix for arbitrary values of q, in the "white" case where the population covariance matrix $\mathbf{C}$ is the identity matrix $\mathbf{C} = \mathbf{1}$.

4.1 Wishart Matrices

4.1.1 Sample Covariance Matrices

We assume that the observed variables x_i^t have zero mean. (Otherwise, we need to remove the sample mean $T^{-1}\sum_t x_i^t$ from x_i^t for each i. For simplicity, we will not consider this case.) Then the sample covariances of the data are given by

$$E_{ij} = \frac{1}{T}\sum_{t=1}^{T} x_i^t x_j^t. \tag{4.1}$$

Thus E_{ij} form an $N \times N$ matrix $\mathbf{E}$, called the *sample covariance matrix* (SCM), which we we write in a compact form as

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

$$\mathbf{E} = \frac{1}{T}\mathbf{H}\mathbf{H}^T, \tag{4.2}$$

where $\mathbf{H}$ is an $N \times T$ data matrix with entries $H_{it} = x_i^t$.

The matrix $\mathbf{E}$ is symmetric and positive semi-definite:

$$\mathbf{E} = \mathbf{E}^T, \quad \text{and} \quad \mathbf{v}^T\mathbf{E}\mathbf{v} = (1/T)\|\mathbf{H}^T\mathbf{v}\|^2 \geq 0, \tag{4.3}$$

for any $\mathbf{v} \in \mathbb{R}^N$. Thus $\mathbf{E}$ is diagonalizable and has all eigenvalues $\lambda_k^{\mathbf{E}} \geq 0$.

We can define another covariance matrix by transposing the data matrix $\mathbf{H}$:

$$\mathbf{F} = \frac{1}{N}\mathbf{H}^T\mathbf{H}. \tag{4.4}$$

The matrix $\mathbf{F}$ is a $T \times T$ matrix, it is also symmetric and positive semi-definite. If the index i $(1 < i < N)$ labels the variables and the index t $(1 < t < T)$ the observations, we can call the matrix $\mathbf{F}$ the covariance of the observations (as opposed to $\mathbf{E}$ the covariance of the variables). F_{ts} measures how similar the observations at t are to those at s – in the above example of neurons, it would measure how similar is the firing pattern at time t and at time s.

As we saw in Section 1.1.3, the matrices $T\mathbf{E}$ and $N\mathbf{F}$ have the same non-zero eigenvalues. Also the matrix $\mathbf{E}$ has at least $N - T$ zero eigenvalues if $N > T$ (and $\mathbf{F}$ has at least $T - N$ zero eigenvalues if $T > N$).

Assume for a moment that $N \leq T$ (i.e. $q \leq 1$), then we know that $\mathbf{F}$ has N (zero or non-zero) eigenvalues inherited from $\mathbf{E}$ and equal to $q^{-1}\lambda_k^{\mathbf{E}}$, and $T - N$ zero eigenvalues. This allows us to write an exact relation between the Stieltjes transforms of $\mathbf{E}$ and $\mathbf{F}$:

$$\begin{aligned} g_T^{\mathbf{F}}(z) &= \frac{1}{T}\sum_{k=1}^{T}\frac{1}{z - \lambda_k^{\mathbf{F}}} \\ &= \frac{1}{T}\left(\sum_{k=1}^{N}\frac{1}{z - q^{-1}\lambda_k^{\mathbf{E}}} + (T - N)\frac{1}{z - 0}\right) \\ &= q^2 g_N^{\mathbf{E}}(qz) + \frac{1-q}{z}. \end{aligned} \tag{4.5}$$

A similar argument with $T < N$ leads to the same Eq. (4.5) so it is actually valid for any value of q. The relationship should be true as well in the large N limit:

$$\mathfrak{g}_{\mathbf{F}}(z) = q^2\mathfrak{g}_{\mathbf{E}}(qz) + \frac{1-q}{z}. \tag{4.6}$$

4.1.2 First and Second Moments of a Wishart Matrix

We now study the SCM $\mathbf{E}$. Assume that the column vectors of $\mathbf{H}$ are drawn independently from a multivariate Gaussian distribution with mean zero and "true" (or "population") covariance matrix $\mathbf{C}$, i.e.

$$\mathbb{E}[H_{it}H_{js}] = C_{ij}\delta_{ts}, \tag{4.7}$$

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

with, again,

$$\mathbf{E} = \frac{1}{T}\mathbf{H}\mathbf{H}^T. \tag{4.8}$$

Sample covariance matrices of this type were first studied by the Scottish mathematician John Wishart (1898–1956) and are now called Wishart matrices.

Recall that if $(X_1, \ldots, X_{2n})$ is a zero-mean multivariate normal random vector, then by Wick's theorem,

$$\mathbb{E}[X_1 X_2 \cdots X_{2n}] = \sum_{\text{pairings}} \prod_{\text{pairs}} \mathbb{E}[X_i X_j] = \sum_{\text{pairings}} \prod_{\text{pairs}} \text{Cov}(X_i, X_j), \tag{4.9}$$

where $\sum_{\text{pairings}} \prod_{\text{pairs}}$ means that we sum over all distinct pairings of $\{X_1, \ldots, X_{2n}\}$ and each summand is the product of the n pairs.

First taking expectation, we obtain that

$$\mathbb{E}[E_{ij}] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} H_{it}H_{jt}\right] = \frac{1}{T}\sum_{t=1}^{T} C_{ij} = C_{ij}. \tag{4.10}$$

Thus, we have $\mathbb{E}[\mathbf{E}] = \mathbf{C}$: as it is well known, the SCM is an unbiased estimator of the true covariance matrix (at least when $\mathbb{E}[x_i^t] = 0$).

For the fluctuations, we need to study the higher order moments of $\mathbf{E}$. The second moment can be calculated as

$$\tau(\mathbf{E}^2) := \frac{1}{NT^2}\mathbb{E}\left[\text{Tr}(\mathbf{H}\mathbf{H}^T\mathbf{H}\mathbf{H}^T)\right] = \frac{1}{NT^2}\sum_{i,j,t,s}\mathbb{E}\left[H_{it}H_{jt}H_{js}H_{is}\right]. \tag{4.11}$$

Then by Wick's theorem, we have (see Fig. 4.1)

$$\begin{aligned} \tau(\mathbf{E}^2) &= \frac{1}{NT^2}\sum_{t,s}\sum_{i,j} C_{ij}^2 + \frac{1}{NT^2}\sum_{t,s}\sum_{i,j} C_{ii}C_{jj}\delta_{ts} + \frac{1}{NT^2}\sum_{t,s}\sum_{i,j} C_{ij}^2\delta_{ts} \\ &= \tau(\mathbf{C}^2) + \frac{N}{T}\tau(\mathbf{C})^2 + \frac{1}{T}\tau(\mathbf{C}^2). \end{aligned} \tag{4.12}$$

Suppose $N, T \to \infty$ with some fixed ratio $N/T = q$ for some constant $q > 0$. The last term on the right hand side then tends to zero and we get

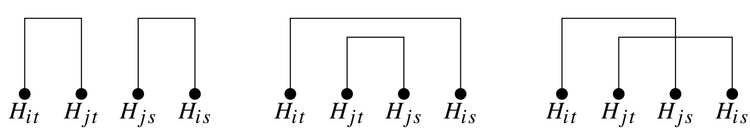

Figure 4.1 Graphical representation of the three Wick's contractions corresponding to the three terms in Eq. (4.12).

$$\tau(\mathbf{E}^2) \to \tau(\mathbf{C}^2) + q\tau(\mathbf{C})^2. \tag{4.13}$$

The variance of the SCM is greater than that of the true covariance by a term proportional to q. When $q \to 0$ we recover perfect estimation and the two matrices have the same variance. If $\mathbf{C} = \alpha\mathbf{1}$ (a multiple of the identity) then $\tau(\mathbf{C}^2) - \tau(\mathbf{C})^2 = 0$ but $\tau(\mathbf{E}^2) - \tau(\mathbf{E})^2 \to q\alpha^2$.

4.1.3 The Law of Wishart Matrices

Next, we give the joint distribution of elements of $\mathbf{E}$. For each fixed column of $\mathbf{H}$, the joint distribution of the elements is

$$P\left(\{H_{it}\}_{i=1}^N\right) = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{C}}} \exp\left[-\frac{1}{2}\sum_{i,j} H_{it}(\mathbf{C})_{ij}^{-1} H_{jt}\right]. \tag{4.14}$$

Taking the product over $1 \le t \le T$ (since the columns are independent), we obtain

$$\begin{aligned} P(\mathbf{H}) &= \frac{1}{(2\pi)^{\frac{NT}{2}} \det \mathbf{C}^{T/2}} \exp\left[-\frac{1}{2}\operatorname{Tr}\left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)\right] \\ &= \frac{1}{(2\pi)^{\frac{NT}{2}} \det \mathbf{C}^{T/2}} \exp\left[-\frac{T}{2}\operatorname{Tr}\left(\mathbf{E}\mathbf{C}^{-1}\right)\right]. \end{aligned} \tag{4.15}$$

Let us now make a change in variables $\mathbf{H} \to \mathbf{E}$. As shown in the technical paragraph 4.1.4, the Jacobian of the transformation is proportional to $(\det \mathbf{E})^{\frac{T-N-1}{2}}$. The following exact expression for the law of the matrix elements was obtained by Wishart:[1]

$$P(\mathbf{E}) = \frac{(T/2)^{NT/2}}{\Gamma_N(T/2)} \frac{(\det \mathbf{E})^{(T-N-1)/2}}{(\det \mathbf{C})^{T/2}} \exp\left[-\frac{T}{2}\operatorname{Tr}\left(\mathbf{E}\mathbf{C}^{-1}\right)\right], \tag{4.16}$$

where Γ_N is the multivariate gamma function. Note that the density is restricted to positive semi-definite matrices $\mathbf{E}$. The Wishart distribution can be thought of as the matrix generalization of the gamma distribution. Indeed for $N = 1$, $P(\mathbf{E})$ reduces to a such a distribution:

$$P_\gamma(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \tag{4.17}$$

where $b = T/(2C)$ and $a = T/2$. Using the identity $\det \mathbf{E} = \exp(\operatorname{Tr}\log \mathbf{E})$, we can rewrite the above expression as[2]

[1] Note that the Wishart distribution is often given with the normalization $\mathbb{E}[\mathbf{E}] = T\mathbf{C}$ as opposed to $\mathbb{E}[\mathbf{E}] = \mathbf{C}$ used here.

[2] Complex and quaternionic Hermitian white Wishart matrices have a similar law of the elements with a factor of β in the exponential:

$$P(\mathbf{W}) \propto \exp\left[-\frac{\beta N}{2}\operatorname{Tr} V(\mathbf{W})\right] \text{ with } V(x) = \frac{N-T-1+2/\beta}{N}\log x + \frac{T}{N}x, \tag{4.18}$$

with β equal to 1, 2 or 4 as usual. The large N limit $V(x)$ is the same in all three cases and is given by Eq. (4.21).

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

$$P(\mathbf{E}) = \frac{(T/2)^{NT/2}}{\Gamma_N(T/2)} \frac{1}{(\det \mathbf{C})^{T/2}} \exp\left[-\frac{T}{2}\operatorname{Tr}\left(\mathbf{E}\mathbf{C}^{-1}\right) + \frac{T-N-1}{2}\operatorname{Tr}\log \mathbf{E}\right]. \quad (4.19)$$

We will denote by $\mathbf{W}$ a SCM with $\mathbf{C} = \mathbf{1}$ and call such a matrix a *white Wishart* matrix. In this case, as $N, T \to \infty$ with $q := N/T$, we get that

$$P(\mathbf{W}) \propto \exp\left[-\frac{N}{2}\operatorname{Tr} V(\mathbf{W})\right], \quad (4.20)$$

where

$$V(\mathbf{W}) := (1 - q^{-1})\log \mathbf{W} + q^{-1}\mathbf{W}. \quad (4.21)$$

Note that the above $P(\mathbf{W})$ is rotationally invariant in the white case. In fact, if a vector $\mathbf{v}$ has Gaussian distribution $\mathcal{N}(0, \mathbf{1}_{N\times N})$, then $\mathbf{O}\mathbf{v}$ has the same distribution $\mathcal{N}(0, \mathbf{1}_{N\times N})$ for any orthogonal matrix $\mathbf{O}$. Hence $\mathbf{O}\mathbf{H}$ has the same distribution as $\mathbf{H}$, which shows that $\mathbf{O}\mathbf{E}\mathbf{O}^T$ has the same distribution as $\mathbf{E}$.

4.1.4 Jacobian of the Transformation $\mathbf{H} \to \mathbf{E}$

The aim here is to compute the volume $\Upsilon(\mathbf{E})$ corresponding to all $\mathbf{H}$'s such that $\mathbf{E} = T^{-1}\mathbf{H}\mathbf{H}^T$:

$$\Upsilon(\mathbf{E}) = \int \mathrm{d}\mathbf{H}\, \delta(\mathbf{E} - T^{-1}\mathbf{H}\mathbf{H}^T). \quad (4.22)$$

Note that this volume is the inverse of the Jacobian of the transformation $\mathbf{H} \to \mathbf{E}$. Next note that one can choose $\mathbf{E}$ to be diagonal, because one can always rotate the integral over $\mathbf{H}$ to an integral over $\mathbf{O}\mathbf{H}$, where $\mathbf{O}$ is the rotation matrix that makes $\mathbf{E}$ diagonal. Now, introducing the Fourier representation of the δ function for all $N(N+1)/2$ independent components of $\mathbf{E}$, one has

$$\Upsilon(\mathbf{E}) = \int \mathrm{d}\mathbf{H}\mathrm{d}\mathbf{A}\exp\left(\mathrm{i}\operatorname{Tr}(\mathbf{A}\mathbf{E} - T^{-1}\mathbf{A}\mathbf{H}\mathbf{H}^T)\right), \quad (4.23)$$

where $\mathbf{A}$ is the symmetric matrix of the corresponding Fourier variables, to which we add a small imaginary part proportional to $\mathbf{1}$ to make all the following integrals well defined. The Gaussian integral over $\mathbf{H}$ can now be performed explicitly for all $t = 1, \ldots, T$, leading to

$$\int \mathrm{d}\mathbf{H}\exp\left(-\mathrm{i}T^{-1}\operatorname{Tr}(\mathbf{A}\mathbf{H}\mathbf{H}^T)\right) \propto (\det \mathbf{A})^{-T/2}, \quad (4.24)$$

leaving us with

$$\Upsilon(\mathbf{E}) \propto \int \mathrm{d}\mathbf{A}\exp(\mathrm{i}\operatorname{Tr}(\mathbf{A}\mathbf{E}))\,(\det \mathbf{A})^{-T/2}. \quad (4.25)$$

We can change variables from $\mathbf{A}$ to $\mathbf{B} = \mathbf{E}^{\frac{1}{2}}\mathbf{A}\mathbf{E}^{\frac{1}{2}}$. The Jacobian of this transformation is

$$\begin{aligned}\prod_i \mathrm{d}\mathbf{A}_{ii}\prod_{j>i}\mathrm{d}\mathbf{A}_{ij} &= \prod_i \mathbf{E}_{ii}^{-1}\prod_{j>i}(\mathbf{E}_{ii}\mathbf{E}_{jj})^{-\frac{1}{2}}\prod_i \mathrm{d}\mathbf{B}_{ii}\prod_{j>i}\mathrm{d}\mathbf{B}_{ij} \\ &= (\det(\mathbf{E}))^{-\frac{N+1}{2}}\prod_i \mathrm{d}\mathbf{B}_{ii}\prod_{j>i}\mathrm{d}\mathbf{B}_{ij}. \end{aligned} \quad (4.26)$$

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

So finally,

$$\Upsilon(\mathbf{E}) \propto \left[\int \mathrm{d}\mathbf{B} \exp\left(\mathrm{i}\,\mathrm{Tr}(\mathbf{B})\right) (\det \mathbf{B})^{-T/2}\right] (\det(\mathbf{E}))^{\frac{T-N-1}{2}}, \tag{4.27}$$

as announced in the main text.

4.2 Marčenko–Pastur Using the Cavity Method

4.2.1 Self-Consistent Equation for the Resolvent

We first derive the asymptotic distribution of eigenvalues of the Wishart matrix with $\mathbf{C} = \mathbf{1}$, i.e. the Marčenko–Pastur distribution. We will use the same method as in the derivation of the Wigner semi-circle law in Section 2.3. In the case $\mathbf{C} = \mathbf{1}$, the $N \times T$ matrix $\mathbf{H}$ is filled with IID standard Gaussian random numbers and we have $\mathbf{W} = (1/T)\mathbf{H}\mathbf{H}^T$.

As in Section 2.3, we wish to derive a self-consistent equation satisfied by the Stieltjes transform:

$$\mathfrak{g}_{\mathbf{W}}(z) = \tau\left(\mathbf{G}_{\mathbf{W}}(z)\right), \quad \mathbf{G}_{\mathbf{W}}(z) := (z\mathbf{1} - \mathbf{W})^{-1}. \tag{4.28}$$

We fix a large N and first write an equation for the element 11 of $\mathbf{G}_{\mathbf{W}}(z)$. We will argue later that $\mathbf{G}_{11}(z)$ converges to $\mathfrak{g}(z)$ with negligible fluctuations. (We henceforth drop the subscript $\mathbf{W}$ as this entire section deals with the white Wishart case.)

Using again the Schur complement formula (1.32), we have that

$$\frac{1}{(\mathbf{G}(z))_{11}} = \mathbf{M}_{11} - \mathbf{M}_{12}(\mathbf{M}_{22})^{-1}\mathbf{M}_{21}, \tag{4.29}$$

where $\mathbf{M} := z\mathbf{1} - \mathbf{W}$, and the submatrices of size, respectively, $[\mathbf{M}_{11}] = 1 \times 1$, $[\mathbf{M}_{12}] = 1 \times (N-1)$, $[\mathbf{M}_{21}] = (N-1) \times 1$, $[\mathbf{M}_{22}] = (N-1) \times (N-1)$. We can expand the above expression and write

$$\frac{1}{(\mathbf{G}(z))_{11}} = z - \mathbf{W}_{11} - \frac{1}{T^2}\sum_{t,s=1}^{T}\sum_{j,k=2}^{N} \mathbf{H}_{1t}\mathbf{H}_{jt}(\mathbf{M}_{22})^{-1}_{jk}\mathbf{H}_{ks}\mathbf{H}_{1s}. \tag{4.30}$$

Note that the three matrices $\mathbf{M}_{22}$, $\mathbf{H}_{jt}$ $(j \geq 2)$ and $\mathbf{H}_{ks}$ $(k \geq 2)$ are independent of the entries H_{1t} for all t. We can write the last term on the right hand side as

$$\frac{1}{T}\sum_{t,s=1}^{N} \mathbf{H}_{1t}\Omega_{ts}\mathbf{H}_{1s} \quad \text{with} \quad \Omega_{ts} := \frac{1}{T}\sum_{j,k=2}^{N} \mathbf{H}_{jt}(\mathbf{M}_{22})^{-1}_{jk}\mathbf{H}_{ks}. \tag{4.31}$$

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

Provided $\gamma^2 := T^{-1} \operatorname{Tr} \Omega^2$ converges to a finite limit when $T \to \infty$,[3] one readily shows that the above sum converges to $T^{-1} \operatorname{Tr} \Omega$ with fluctuations of the order of $\gamma T^{-\frac{1}{2}}$. So we have, for large T,

$$
\begin{aligned}
\frac{1}{(\mathbf{G}(z))_{11}} &= z - \mathbf{W}_{11} - \frac{1}{T} \sum_{2 \le j,k \le N} \frac{\sum_t H_{kt} H_{jt}}{T} (\mathbf{M}_{22})^{-1}_{jk} + O\left(T^{-\frac{1}{2}}\right) \\
&= z - \mathbf{W}_{11} - \frac{1}{T} \sum_{2 \le j,k \le N} \mathbf{W}_{kj} (\mathbf{M}_{22})^{-1}_{jk} + O\left(T^{-\frac{1}{2}}\right) \\
&= z - 1 - \frac{1}{T} \operatorname{Tr} \mathbf{W}_2 \mathbf{G}_2(z) + O\left(T^{-\frac{1}{2}}\right),
\end{aligned} \tag{4.32}
$$

where in the last step we have used the fact that $\mathbf{W}_{11} = 1 + O(T^{-\frac{1}{2}})$ and noted $\mathbf{W}_2$ and $\mathbf{G}_2(z)$ the SCM and resolvent of the $N-1$ variables excluding (1). We can rewrite the trace term:

$$
\begin{aligned}
\operatorname{Tr}(\mathbf{W}_2 \mathbf{G}_2(z)) &= \operatorname{Tr}\left(\mathbf{W}_2 (z\mathbf{1} - \mathbf{W}_2)^{-1}\right) \\
&= -\operatorname{Tr} \mathbf{1} + z \operatorname{Tr}\left((z\mathbf{1} - \mathbf{W}_2)^{-1}\right) \\
&= -\operatorname{Tr} \mathbf{1} + z \operatorname{Tr} \mathbf{G}_2(z).
\end{aligned} \tag{4.33}
$$

In the region where $\operatorname{Tr} \mathbf{G}(z)/N$ converges for large N to the deterministic $\mathfrak{g}(z)$, $\operatorname{Tr} \mathbf{G}_2(z)/N$ should also converge to the same limit as $\mathbf{G}_2(z)$ is just an $(N-1) \times (N-1)$ version of $\mathbf{G}(z)$. So in the region of convergence we have

$$
\frac{1}{(\mathbf{G}(z))_{11}} = z - 1 + q - qz\mathfrak{g}(z) + O\left(N^{-\frac{1}{2}}\right), \tag{4.34}
$$

where we have introduced $q = N/T = O(1)$, such that $N^{-\frac{1}{2}}$ and $T^{-\frac{1}{2}}$ are of the same order of magnitude. This last equation states that $1/\mathbf{G}_{11}(z)$ has negligible fluctuations and can safely be replaced by its expectation value, i.e.

$$
\begin{aligned}
\frac{1}{(\mathbf{G}(z))_{11}} &= \mathbb{E}\left[\frac{1}{(\mathbf{G}(z))_{11}}\right] + O\left(N^{-\frac{1}{2}}\right) \\
&= \frac{1}{\mathbb{E}\left[(\mathbf{G}(z))_{11}\right]} + O\left(N^{-\frac{1}{2}}\right).
\end{aligned} \tag{4.35}
$$

By rotational invariance of $\mathbf{W}$, we have

$$
\mathbb{E}[\mathbf{G}(z)_{11}] = \frac{1}{N} \mathbb{E}[\operatorname{Tr}(\mathbf{G}(z))] \to \mathfrak{g}(z). \tag{4.36}
$$

[3] It can be self-consistently checked from the solution below that $\lim_{T\to\infty} \gamma^2 = -q\mathfrak{g}'_{\mathbf{W}}(z)$.

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

In the large N limit we obtain the following self-consistent equation for $\mathfrak{g}(z)$:

$$\frac{1}{\mathfrak{g}(z)} = z - 1 + q - qz\mathfrak{g}(z). \tag{4.37}$$

4.2.2 Solution and Density of Eigenvalues

Solving (4.37) we obtain

$$\mathfrak{g}(z) = \frac{z + q - 1 \pm \sqrt{(z + q - 1)^2 - 4qz}}{2qz}. \tag{4.38}$$

The argument of the square-root is quadratic in z and its roots (the edge of spectrum) are given by

$$\lambda_\pm = (1 \pm \sqrt{q})^2. \tag{4.39}$$

Finding the correct branch is quite subtle, this will be the subject of Section 4.2.3. We will see that the form

$$\mathfrak{g}(z) = \frac{z - (1 - q) - \sqrt{z - \lambda_+}\sqrt{z - \lambda_-}}{2qz} \tag{4.40}$$

has all the correct analytical properties. Note that for $z = x - i\eta$ with $x \neq 0$ and $\eta \to 0$, $g(z)$ can only have an imaginary part if $\sqrt{(x - \lambda_+)(x - \lambda_-)}$ is imaginary. Then using (2.47), we get the famous Marčenko–Pastur distribution for the bulk:

$$\rho(x) = \frac{1}{\pi} \lim_{\eta \to 0+} \operatorname{Im} \mathfrak{g}(x - i\eta) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi qx}, \quad \lambda_- < x < \lambda_+. \tag{4.41}$$

Moreover, by studying the behavior of Eq. (4.40) near $z = 0$ one sees that there is a pole at 0 when $q > 1$. This gives a delta mass as $z \to 0$:

$$\frac{q - 1}{q}\delta(x), \tag{4.42}$$

which corresponds to the $N - T$ trivial zero eigenvalues of E in the $N > T$ case. Combining the above discussions, the full Marčenko–Pastur law can be written as

$$\rho_{\mathrm{MP}}(x) = \frac{\sqrt{\left[(\lambda_+ - x)(x - \lambda_-)\right]_+}}{2\pi qx} + \frac{q - 1}{q}\delta(x)\Theta(q - 1), \tag{4.43}$$

where we denote $[a]_+ := \max\{a, 0\}$ for any $a \in \mathbb{R}$, and

$$\Theta(q - 1) := \begin{cases} 0, & \text{if } q \leq 1, \\ 1 & \text{if } q > 1. \end{cases} \tag{4.44}$$

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

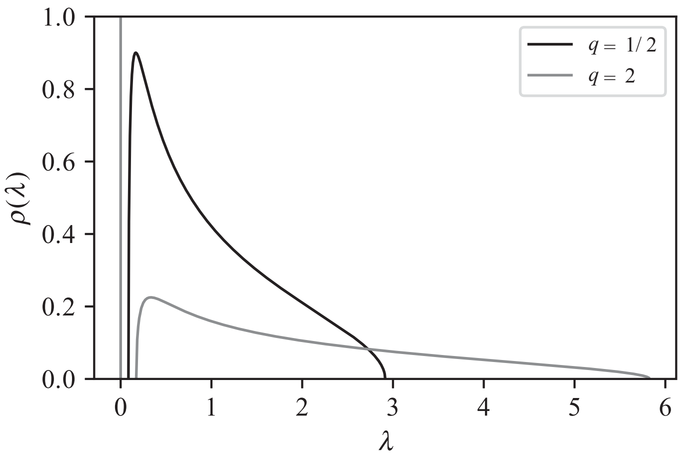

Figure 4.2 Marčenko–Pastur distribution: density of eigenvalues for a Wishart matrix for $q = 1/2$ and $q = 2$. Note that for $q = 2$ there is a Dirac mass at zero ($\frac{1}{2}\delta(\lambda)$). Also note that the two bulk densities are the same up to a rescaling and normalization $\rho_{1/q}(\lambda) = q^2 \rho_q(q\lambda)$.

Note that the Stieltjes transforms (Eq. (4.40)) for q and $1/q$ are related by Eq. (4.5). As a consequence the bulk densities for q and $1/q$ are the same when properly rescaled (see Fig. 4.2):

$$\rho_{1/q}(\lambda) = q^2 \rho_q(q\lambda). \tag{4.45}$$

Exercise 4.2.1 Properties of the Marčenko–Pastur solution

We saw that the Stieltjes transform of a large Wishart matrix (with $q = N/T$) should be given by

$$g(z) = \frac{z + q - 1 \pm \sqrt{(z+q-1)^2 - 4qz}}{2qz}, \tag{4.46}$$

where the sign of the square-root should be chosen such that $g(z) \to 1/z$ when $z \to \pm\infty$.

(a) Show that the zeros of the argument of the square-root are given by $\lambda_\pm = (1 \pm \sqrt{q})^2$.

(b) The function

$$g(z) = \frac{z + q - 1 - \sqrt{(z - \lambda_-)}\sqrt{(z - \lambda_+)}}{2qz} \tag{4.47}$$

should have the right properties. Show that it behaves as $g(z) \to 1/z$ when $z \to \pm\infty$. By expanding in powers of $1/z$ up to $1/z^3$ compute the first and second moments of the Wishart distribution.

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

(c) Show that Eq. (4.47) is regular at $z = 0$ when $q < 1$. In that case, compute the first inverse moment of the Wishart matrix $\tau(\mathbf{E}^{-1})$. What happens when $q \to 1$? Show that Eq. (4.47) has a pole at $z = 0$ when $q > 1$ and compute the value of this pole.

(d) The non-zero eigenvalues should be distributed according to the Marčenko–Pastur distribution:

$$\rho_q(x) = \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2\pi q x}. \tag{4.48}$$

Show that this distribution is correctly normalized when $q < 1$ but not when $q > 1$. Use what you know about the pole at $z = 0$ in that case to correctly write down $\rho_q(x)$ when $q > 1$.

(e) In the case $q = 1$, Eq. (4.48) has an integrable singularity at $x = 0$. Write a simpler formula for $\rho_1(x)$. Let u be the square of an eigenvalue from a Wigner matrix of unit variance, i.e. $u = y^2$ where y is distributed according to the semi-circular law $\rho(y) = \sqrt{4-y^2}/(2\pi)$. Show that u is distributed according to $\rho_1(x)$. This result is *a priori* not obvious as a Wigner matrix is symmetric while the square matrix $\mathbf{H}$ is generally not; nevertheless, moments of high-dimensional matrices of the form $\mathbf{H}\mathbf{H}^T$ are the same whether the matrix $\mathbf{H}$ is symmetric or not.

(f) Generate three matrices $\mathbf{E} = \mathbf{H}\mathbf{H}^T/T$ where the matrix $\mathbf{H}$ is an $N \times T$ matrix of IID Gaussian numbers of variance 1. Choose a large N and three values of T such that $q = N/T$ equals $\{1/2, 1, 2\}$. Plot a normalized histogram of the eigenvalues in the three cases vs the corresponding Marčenko–Pastur distribution; don't show the peak at zero. In the case $q = 2$, how many zero eigenvalues do you expect? How many do you get?

4.2.3 The Correct Root of the Stieltjes Transform

In our study of random matrices we will often encounter limiting Stieltjes transforms that are determined by quadratic or higher order polynomial equations, and the problem of choosing the correct solution (or branch) will come up repeatedly.

Let us go back to the unit Wigner matrix case where we found (see Section 2.3.2)

$$\mathfrak{g}(z) = \frac{z \pm \sqrt{z^2-4}}{2}. \tag{4.49}$$

On the one hand we want $\mathfrak{g}(z)$ that behaves like $1/z$ as $|z| \to \infty$ and we want the solution to be analytical everywhere but on the real axis in $[-2, 2]$. The square-root term must thus behave as $-z$ for real z when $z \to \pm\infty$. The standard definition of the square-root behaves as $\sqrt{z^2} \sim |z|$ and cannot be made to have the correct sign on both sides. Another issue with $\sqrt{z^2-4}$ is that it has a more extended branch cut than allowed. We expect the

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

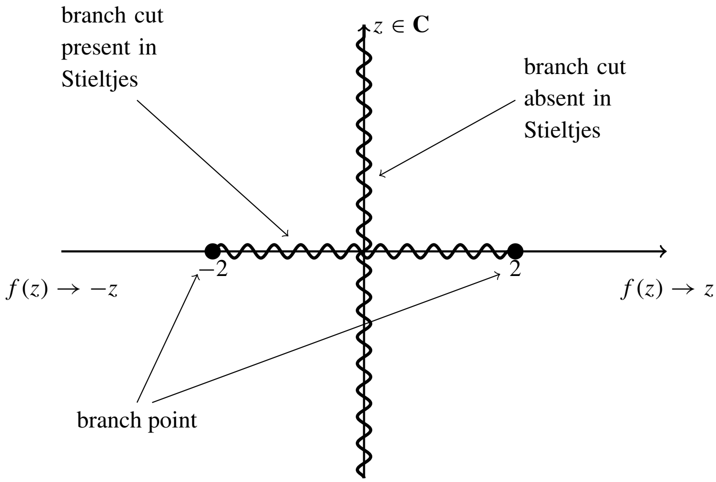

Figure 4.3 The branch cuts of $f(z) = \sqrt{z^2 - 4}$. The vertical branch cut (for z pure imaginary) should not be present in the Stieltjes transform of the Wigner ensemble. We have $f(\pm 0^+ + \mathrm{i}x) \approx \pm\mathrm{i}\sqrt{x^2 + 4}$; this branch cut can be eliminated by multiplying $f(z)$ by sign(Re z).

function $\mathfrak{g}(z)$ to be analytic everywhere except for real $z \in [-2, 2]$. The branch cut of a standard square-root is a set of points where its argument is real and negative. In the case of $\sqrt{z^2 - 4}$, this includes the interval $[-2, 2]$ as expected but also the pure imaginary line $z = \mathrm{i}x$ (Fig. 4.3). The finite N Stieltjes transform is perfectly regular on the imaginary axis so we expect its large N to be regular there as well.

For the unit Wigner matrix, there are at least three solutions to the branch problem:

$$\mathfrak{g}_1(z) = \frac{z - z\sqrt{1 - \frac{4}{z^2}}}{2}, \tag{4.50}$$

$$\mathfrak{g}_2(z) = \frac{z - \sqrt{z-2}\,\sqrt{z+2}}{2}, \tag{4.51}$$

$$\mathfrak{g}_3(z) = \frac{z - \mathrm{sign}(\mathrm{Re}\, z)\sqrt{z^2 - 4}}{2}. \tag{4.52}$$

All three definitions behave as $\mathfrak{g}(z) \sim 1/z$ at infinity. For the second one, we need to define the square-root of a negative real number. If we define it as $\mathrm{i}\sqrt{|z|}$, the two factors of i give a -1 for real $z < -2$. The three functions also have the correct branch cuts. For the first one, one can show that the argument of the square-root can be a negative real number only if $z \in (-2, 2)$, there are no branch cuts elsewhere in the complex plane. For the second one, there seems to be a branch cut for all real $z < 2$, but a closer inspection reveals that around real $z < -2$ the function has no discontinuity as one goes up and down the imaginary axis, as the two branch cuts exactly compensate each other. For the third one, the discontinuous sign function exactly cancels the branch cut on the pure imaginary axis (Fig. 4.3).

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

For z with a large positive real part the three functions are clearly the same. Since they are analytic functions everywhere except on the same branch cuts, they are the same and unique function $\mathfrak{g}(z)$.

For a Wigner matrix shifted by λ_0 and of variance σ^2 we can scale and shift the eigenvalues, now equal $\lambda_\pm = \lambda_0 \pm 2\sigma$ and find

$$\mathfrak{g}_1(z) = \frac{z - \lambda_0 - (z - \lambda_0)\sqrt{1 - \frac{4\sigma^2}{(z-\lambda_0)^2}}}{2\sigma}, \tag{4.53}$$

$$\mathfrak{g}_2(z) = \frac{z - \lambda_0 - \sqrt{z - \lambda_+}\sqrt{z - \lambda_-}}{2\sigma}, \tag{4.54}$$

$$\mathfrak{g}_3(z) = \frac{z - \lambda_0 - \operatorname{sign}(\operatorname{Re} z - \lambda_0)\sqrt{(z - \lambda_0)^2 - 4\sigma^2}}{2\sigma}. \tag{4.55}$$

The three definitions are still equivalent as they are just the result of a shift and a scaling of the same function.

For more complicated problems, writing explicitly any one of these three prescriptions can quickly become very cumbersome (except maybe in cases where $\lambda_+ + \lambda_- = 0$). We propose here a new notation. When finding the correct square-root of a second degree polynomial we will write

$$\begin{aligned} \sqrt[\oplus]{az^2 + bz + c} &:= \sqrt{a}\sqrt{z - \lambda_+}\sqrt{z - \lambda_-} \\ &= \sqrt{a}(z - \lambda_0)\sqrt{1 - \frac{\Delta}{(z - \lambda_0)^2}} \\ &= \operatorname{sign}(\operatorname{Re} z - \lambda_0)\sqrt{az^2 + bz + c}, \end{aligned} \tag{4.56}$$

for $a > 0$ and where $\lambda_\pm = \lambda_0 \pm \sqrt{\Delta}$ are the roots of $az^2 + bz + c$ assumed to be real. While the notation is defined everywhere in the complex plane, it is easily evaluated for real arguments:

$$\sqrt[\oplus]{ax^2 + bx + c} = \begin{cases} -\sqrt{ax^2 + bx + c} & \text{for } x \le \lambda_-, \\ \sqrt{ax^2 + bx + c} & \text{for } x \ge \lambda_+. \end{cases} \tag{4.57}$$

The value on the branch cut is ill-defined but we have

$$\lim_{z \to x - \mathrm{i}0^+} \sqrt[\oplus]{az^2 + bz + c} = \mathrm{i}\sqrt{|ax^2 + bx + c|} \quad \text{for } \lambda_- < x < \lambda_+. \tag{4.58}$$

With our new notation, we can now safely write for the white Wishart:

$$\mathfrak{g}(z) = \frac{z + q - 1 - \sqrt[\oplus]{(z + q - 1)^2 - 4qz}}{2qz}, \tag{4.59}$$

or, more explicitly, using the second prescription:

$$\mathfrak{g}(z) = \frac{z + q - 1 - \sqrt{z - \lambda_+}\sqrt{z - \lambda_-}}{2qz}, \tag{4.60}$$

where $\lambda_\pm = (1 \pm \sqrt{q})^2$.

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

Exercise 4.2.2 Finding the correct root

(a) For the unit Wigner Stieltjes transform show that regardless of choice of sign in Eq. (4.49) the point $z = 2\text{i}$ is located on a branch cut and the function is discontinuous at that point.

(b) Compute the value of Eqs. (4.50), (4.51) and (4.52) at $z = 2\text{i}$. Hint: for $\mathfrak{g}_2(z)$ write $-2 + 2\text{i} = \sqrt{8}\text{e}^{3\text{i}\pi/4}$ and similarly for $2 + 2\text{i}$. The definition $\mathfrak{g}_3(z)$ is ambiguous for $z = 2\text{i}$, compute the limiting value on both sides: $z = 0^+ + 2\text{i}$ and $z = 0^- + 2\text{i}$.

4.2.4 General (Non-White) Wishart Matrices

Recall our definition of a Wishart matrix from Section 4.1.2: a Wishart matrix is a matrix $\mathbf{E}_\mathbf{C}$ defined as

$$\mathbf{E}_\mathbf{C} = \frac{1}{T}\mathbf{H}_\mathbf{C}\mathbf{H}_\mathbf{C}^T, \tag{4.61}$$

where $\mathbf{H}_\mathbf{C}$ is an $N \times T$ rectangular matrix with independent columns. Each column is a random Gaussian vector with covariance matrix $\mathbf{C}$; $\mathbf{E}_\mathbf{C}$ corresponds to the sample (empirical) covariance matrix of variables characterized by a population (true) covariance matrix $\mathbf{C}$.

To understand the case where the true matrix $\mathbf{C}$ is different from the identity we first discuss how to generate a multivariate Gaussian vector with covariance matrix $\mathbf{C}$. We diagonalize $\mathbf{C}$ as

$$\mathbf{C} = \mathbf{O}\Lambda\mathbf{O}^T, \quad \Lambda = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_N^2 \end{pmatrix}. \tag{4.62}$$

The square-root of $\mathbf{C}$ can be defined as[4]

$$\mathbf{C}^{\frac{1}{2}} = \mathbf{O}\Lambda^{\frac{1}{2}}\mathbf{O}^T, \quad \Lambda^{\frac{1}{2}} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_N \end{pmatrix}. \tag{4.63}$$

We now generate N IID unit Gaussian random variables x_i, $1 \le i \le N$, which form a random column vector $\mathbf{x}$ with entries x_i. Then we can generate the vector $\mathbf{y} = \mathbf{C}^{\frac{1}{2}}\mathbf{x}$. We claim that $\mathbf{y}$ is a multivariate Gaussian vector with covariance matrix $\mathbf{C}$. In fact, $\mathbf{y}$ is a linear combination of multivariate Gaussians, so it must itself be multivariate Gaussian. On the other hand, we have, using $\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \mathbf{1}$,

[4] This is the canonical definition of the square-root of a matrix, but this definition is not unique – see the technical paragraph below.

https://doi.org/10.1017/9781108768900.006 Published online by Cambridge University Press

$$\mathbb{E}(\mathbf{y}\mathbf{y}^T) = \mathbb{E}(\mathbf{C}^{\frac{1}{2}}\mathbf{x}\mathbf{x}^T\mathbf{C}^{\frac{1}{2}}) = \mathbf{C}. \tag{4.64}$$

By repeating the argument above for every column of $\mathbf{H}_\mathbf{C}$, $t = 1, \dots, T$, we see that this matrix can be written as $\mathbf{H}_\mathbf{C} = \mathbf{C}^{\frac{1}{2}}\mathbf{H}$, with $\mathbf{H}$ a rectangular matrix with IID unit Gaussian entries. The matrix $\mathbf{E}_\mathbf{C}$ is then equivalent to

$$\mathbf{E}_\mathbf{C} = \frac{1}{T}\mathbf{H}_\mathbf{C}\mathbf{H}_\mathbf{C}^T = \frac{1}{T}\mathbf{C}^{\frac{1}{2}}\mathbf{H}\mathbf{H}^T\mathbf{C}^{\frac{1}{2}} = \mathbf{C}^{\frac{1}{2}}\mathbf{W}_q\mathbf{C}^{\frac{1}{2}}, \tag{4.65}$$

where $\mathbf{W}_q = \frac{1}{T}\mathbf{H}\mathbf{H}^T$ is a white Wishart matrix with $q = N/T$.

We will see later that the above combination of matrices is called the *free product* of $\mathbf{C}$ and $\mathbf{W}$. Free probability will allow us to compute the resolvent and the spectrum in the case of a general $\mathbf{C}$ matrix.

The variables $\mathbf{x}$ defined above are called the "whitened" version of $\mathbf{y}$. If a zero mean random vector $\mathbf{y}$ has positive definite covariance matrix $\mathbf{C}$, we can define a *whitening* of $\mathbf{y}$ as a linear combination $\mathbf{x} = \mathbf{M}\mathbf{y}$ such that $\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \mathbf{1}$. One can show that the matrix $\mathbf{M}$ satisfies $\mathbf{M}^T\mathbf{M} = \mathbf{C}^{-1}$ and has to be of the form $\mathbf{M} = \mathbf{O}\mathbf{C}^{-\frac{1}{2}}$, where $\mathbf{O}$ can be any orthogonal matrix and $\mathbf{C}^{\frac{1}{2}}$ the symmetric square-root of $\mathbf{C}$ defined above. Since $\mathbf{O}$ is arbitrary the procedure is not unique, which leads to three interesting choices for whitened varaibles:

- Perhaps the most natural one is the symmetric or *Mahalanobis* whitening where $\mathbf{M} = \mathbf{C}^{-\frac{1}{2}}$. In addition to being the only whitening scheme with a symmetric matrix $\mathbf{M}$, the white variables $\mathbf{x} = \mathbf{C}^{-\frac{1}{2}}\mathbf{y}$ are the closest to $\mathbf{y}$ in the following sense: the distance

$$||\mathbf{x} - \mathbf{y}||_{\mathbf{C}^\alpha} := \mathbb{E}\,\mathrm{Tr}\left[(\mathbf{x} - \mathbf{y})^T\mathbf{C}^\alpha(\mathbf{x} - \mathbf{y})\right] \tag{4.66}$$

  is minimal over all other choices of $\mathbf{O}$ for any α. The case $\alpha = -1$ is called the Mahalanobis norm.
- Triangular or Gram–Schmidt whitening where the vector $\mathbf{x}$ can be constructed using the Gram–Schmidt orthonormalization procedure. If one starts from the bottom with $x_N = y_N/\sqrt{C_{NN}}$, then the matrix $\mathbf{M}$ is upper triangular. The matrix $\mathbf{M}$ can be computed efficiently using the Cholesky decomposition of $\mathbf{C}^{-1}$. The Cholesky decomposition of a symmetric positive definite matrix $\mathbf{A}$ amounts to finding a lower triangular matrix $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^T = \mathbf{A}$. In the present case, $\mathbf{A} = \mathbf{C}^{-1}$ and the matrix $\mathbf{M}$ we are looking for is given by

$$\mathbf{M} = \mathbf{L}^T. \tag{4.67}$$

  This scheme has the advantage that the whitened variable x_k only depends on physical variables y_ℓ for $\ell \geq k$. In finance, for example, this allows one to construct whitened returns of a given stock using only the returns of itself and those of (say) more liquid stocks.
- Eigenvalue (or PCA) whitening where $\mathbf{O}$ corresponds to the eigenbasis of $\mathbf{C}$, i.e. such that $\mathbf{C} = \mathbf{O}\Lambda\mathbf{O}^T$ where Λ is diagonal. $\mathbf{M}$ is then computed as $\mathbf{M} = \Lambda^{-\frac{1}{2}}\mathbf{O}^T$. The whitened variables $\mathbf{x}$ are then called the normalized principal components of $\mathbf{y}$.

Bibliographical Notes

- For a historical perspective on Wishart matrices and the Marčenko–Pastur, see
  - J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1-2):32–52, 1928,
  - V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.
- For more recent material on the content of this chapter, see e.g.
  - L. Pastur and M. Scherbina. *Eigenvalue Distribution of Large Random Matrices*. American Mathematical Society, Providence, Rhode Island, 2010,
  - Z. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer-Verlag, New York, 2010,
  - A. M. Tulino and S. Verdú. *Random Matrix Theory and Wireless Communications*. Now publishers, Hanover, Mass., 2004,
  - R. Couillet and M. Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, Cambridge, 2011,
  - G. Livan, M. Novaes, and P. Vivo. *Introduction to Random Matrices: Theory and Practice*. Springer, New York, 2018,

  and also
  - J. W. Silverstein and S.-I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995,
  - A. Sengupta and P. P. Mitra. Distributions of singular values for some random matrices. *Physical Review E*, 60(3):3389, 1999.
- A historical remark on the Cholesky decomposition: André-Louis Cholesky served in the French military as an artillery officer and was killed in battle a few months before the end of World War I; his discovery was published posthumously by his fellow officer Commandant Benoît in the *Bulletin Géodésique* (Wikipedia).