19

Eigenvector Overlaps and Rotationally Invariant Estimators

19.1 Eigenvector Overlaps

19.1.1 Setting the Stage

We saw in the first two parts of this book how tools from RMT allow one to infer many properties of the eigenvalue distribution, encoded in the trace of the resolvent of the random matrix under scrutiny. As in the previous chapters, random matrices of particular interest are of the form

$$\mathbf{E} = \mathbf{C} + \mathbf{X}, \qquad \text{or} \qquad \mathbf{E} = \mathbf{C}^{\frac{1}{2}}\mathbf{W}\mathbf{C}^{\frac{1}{2}}, \tag{19.1}$$

where $\mathbf{X}$ and $\mathbf{W}$ represent some "noise", for example $\mathbf{X}$ might be a Wigner matrix in the additive case and $\mathbf{W}$ a white Wishart matrix in the multiplicative case, whereas $\mathbf{C}$ is the "true", uncorrupted matrix that one would like to measure. One often calls $\mathbf{E}$ the *sample* matrix and $\mathbf{C}$ the *population* matrix.

In this section we want to discuss the properties of the eigenvectors of $\mathbf{E}$, and in particular their relation with the eigenvectors of $\mathbf{C}$. There are, at least, two natural questions about the eigenvectors of the sample matrix $\mathbf{E}$:

1 How similar are sample eigenvectors $[\mathbf{v}_i]_{i\in(1,N)}$ of $\mathbf{E}$ and the true ones $[\mathbf{u}_i]_{i\in(1,N)}$ of $\mathbf{C}$?
2 What information can we learn by observing two independent realizations – say $\mathbf{E} = \mathbf{C}^{\frac{1}{2}}\mathbf{W}\mathbf{C}^{\frac{1}{2}}$ and $\mathbf{E}' = \mathbf{C}^{\frac{1}{2}}\mathbf{W}'\mathbf{C}^{\frac{1}{2}}$ in the multiplicative case – that remain correlated through $\mathbf{C}$?

A natural quantity to characterize the similarity between two arbitrary vectors – say $\boldsymbol{\chi}$ and $\boldsymbol{\zeta}$ – is the scalar product of $\boldsymbol{\chi}$ and $\boldsymbol{\zeta}$. More formally, we define the "overlap" as $\boldsymbol{\chi}^T\boldsymbol{\zeta}$. Since the eigenvectors of real symmetric matrices are only defined up to a sign, it is in fact more natural to consider the squared overlaps $(\boldsymbol{\chi}^T\boldsymbol{\zeta})^2$. In the first problem alluded to above, we want to understand the relation between the eigenvectors of the sample matrix $[\mathbf{v}_i]_{i\in(1,N)}$ and those of the population matrix $[\mathbf{u}_i]_{i\in(1,N)}$. The matrix of squared overlaps is defined as $(\mathbf{v}_i{}^T\mathbf{u}_j)^2$, which actually forms a so-called bi-stochastic matrix (positive elements with the sums over both rows and columns all equal to unity).

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

In order to study these overlaps, the central tool is again the resolvent matrix (and not its normalized trace as for the Stieltjes transform), which we recall is defined as

$$\mathbf{G}_{\mathbf{A}}(z) = (z\mathbf{1} - \mathbf{A})^{-1}, \tag{19.2}$$

for any arbitrary symmetric matrix $\mathbf{A}$. Now, if we expand $\mathbf{G}_{\mathbf{E}}$ over the eigenvectors $\mathbf{v}$ of $\mathbf{E}$, we obtain that

$$\mathbf{u}^T \mathbf{G}_{\mathbf{E}}(z)\mathbf{u} = \sum_{i=1}^{N} \frac{\left(\mathbf{u}^T \mathbf{v}_i\right)^2}{z - \lambda_i}, \tag{19.3}$$

for any $\mathbf{u}$ in $\mathbb{R}^N$.

We thus see from Eq. (19.3) that each pole of the resolvent defines a projection onto the corresponding sample eigenvectors. This suggests that the techniques we need to apply are very similar to the ones used above to study the density of states. However, one should immediately stress that contrarily to eigenvalues, each eigenvector $\mathbf{v}_i$ for any given i continues to fluctuate when $N \to \infty$ and never reaches a deterministic limit. As a consequence, we will need to introduce some averaging procedure to obtain a well-defined result. We will thus consider the following quantity:

$$\Phi(\lambda_i, \mu_j) \;:=\; N\mathbb{E}[(\mathbf{v}_i^T \mathbf{u}_j)^2], \tag{19.4}$$

where the expectation $\mathbb{E}$ can be interpreted either as an average over different realizations of the randomness or, perhaps more meaningfully for applications, as an average *for a fixed sample* over small intervals of sample eigenvalues, of width $\mathrm{d}\lambda = \eta$. We choose η in the range $1 \gg \eta \gg N^{-1}$ (say $\eta = N^{-1/2}$) such that there are many eigenvalues in the interval $\mathrm{d}\lambda$, while keeping $\mathrm{d}\lambda$ sufficiently small for the spectral density to be approximately constant. Interestingly, the two procedures lead to the same result for large matrices, i.e. the locally smoothed quantity $\Phi(\lambda, \mu)$ is "self-averaging". A way to do this smoothing automatically is, as we explained in Chapter 2, to choose $z = \lambda_i - \mathrm{i}\eta$ in Eq. (19.3), leading to

$$\mathrm{Im}\, \mathbf{u}_j^T \mathbf{G}_{\mathbf{E}}(\lambda_i - \mathrm{i}\eta)\mathbf{u}_j \approx \pi \rho_{\mathbf{E}}(\lambda_i) \times \Phi(\lambda_i, \mu_j), \tag{19.5}$$

provided η is in the range $1 \gg \eta \gg N^{-1}$. Note that we have replaced $N\left(\mathbf{u}_j^T \mathbf{v}_i\right)^2$ with $\Phi(\lambda_i, \mu_j)$, to emphasize the fact that we expect typical square overlaps to be of order $1/N$, such that Φ is of order unity when $N \to \infty$. This assumption will indeed be shown to hold below. In fact, when $\mathbf{u}_j$ and $\mathbf{v}_i$ are completely uncorrelated, one finds $\Phi(\lambda_i, \mu_j) = 1$.

For the second question, the main quantity of interest is, similarly, the (mean squared) overlap between the eigenvectors of two independent noisy matrices $\mathbf{E}$ and $\mathbf{E}'$:

$$\Psi(\lambda_i, \lambda'_j) \;:=\; N\mathbb{E}[(\mathbf{v}_i^T \mathbf{v}'_j)^2], \tag{19.6}$$

where $[\lambda'_i]_{i\in(1,N)}$ and $[\mathbf{v}'_i]_{i\in(1,N)}$ are the eigenvalues and eigenvectors of $\mathbf{E}'$, i.e. another sample matrix that is independent from $\mathbf{E}$ but with the same underlying population

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

matrix $\mathbf{C}$. In order to get access to $\Phi(\lambda_i, \lambda'_j)$ defined in Eq. (19.5), one should consider the following quantity:

$$\psi(z, z') = \frac{1}{N} \operatorname{Tr}\left[\mathbf{G}_{\mathbf{E}}(z)\mathbf{G}_{\mathbf{E}'}(z')\right]. \tag{19.7}$$

After simple manipulations one readily obtains a generalized Sokhotski–Plemelj formula, where η is such that $1 \gg \eta \gg N^{-1}$:

$$\operatorname{Re}\left[\psi(\lambda_i - \mathrm{i}\eta, \lambda'_i + \mathrm{i}\eta) - \psi(\lambda_i - \mathrm{i}\eta, \lambda'_i - \mathrm{i}\eta)\right] \approx 2\pi^2 \rho_{\mathbf{E}}(\lambda_i)\rho_{\mathbf{E}'}(\lambda'_i)\Psi(\lambda_i, \lambda'_j). \tag{19.8}$$

This representation allows one to obtain interesting results for the overlaps between the eigenvectors of two independently drawn random matrices, see Eq. (19.14).

19.1.2 Overlaps in the Additive Case

Now, we can use the subordination relation for the resolvent of the sum of two free matrices established in Chapter 13, Eq. (13.44), which in the present case reads

$$\mathbb{E}[\mathbf{G}_{\mathbf{E}}(z)] = \mathbf{G}_{\mathbf{C}}\left(z - R_{\mathbf{X}}(\mathfrak{g}_{\mathbf{E}}(z))\right). \tag{19.9}$$

Since we choose $\mathbf{u}_j$ to be an eigenvector of $\mathbf{C}$ with eigenvalue μ_j, one finds

$$\mathbf{u}_j^T \mathbf{G}_{\mathbf{E}}(\lambda_i - \mathrm{i}\eta)\mathbf{u}_j = \frac{1}{\lambda_i - \mathrm{i}\eta - R_{\mathbf{X}}\left(\mathfrak{g}_{\mathbf{E}}(\lambda_i - \mathrm{i}\eta)\right) - \mu_j}, \tag{19.10}$$

where we have dropped the expectation value as the left hand side is self-averaging when η is in the correct range. The imaginary part of this quantity, calculated for $\eta \to 0$, gives access to $\Phi(\lambda_i, \mu_j)$. The formula simplifies in the common case where the noise matrix $\mathbf{X}$ is a Wigner matrix, such that $R_{\mathbf{X}}(z) = \sigma^2 z$. In this case, one finally obtains a Lorentzian shape for the squared overlaps:

$$\Phi(\lambda, \mu) = \frac{\sigma^2}{(\mu - \lambda + \sigma^2 \mathfrak{h}_{\mathbf{E}}(\lambda))^2 + \sigma^4 \pi^2 \rho_{\mathbf{E}}(\lambda)^2}, \tag{19.11}$$

where we have decomposed the Stieltjes transform into its real and imaginary parts as $\mathfrak{g}_{\mathbf{E}}(x) = \mathfrak{h}_{\mathbf{E}}(x) + \mathrm{i}\pi\rho_{\mathbf{E}}(x)$; note that $\mathfrak{h}_{\mathbf{E}}(x)$ is equal to π times the Hilbert transform of $\rho_{\mathbf{E}}(x)$.

In Figure 19.1, we illustrate this formula in the case where $\mathbf{C}$ is a Wigner matrix with parameter $\sigma^2 = 1$. For a fixed λ, the overlap peaks for

$$\mu = \lambda - \sigma^2 \mathfrak{h}_{\mathbf{E}}(\lambda), \tag{19.12}$$

with a width $\sim \sigma^2 \rho_{\mathbf{E}}(\lambda)$. When $\sigma \to 0$, i.e. in the absence of noise, one recovers

$$\Phi(\lambda, \mu) \to \delta(\lambda - \mu), \tag{19.13}$$

as expected since in this case the eigenvectors of $\mathbf{E}$ are trivially the same as those of $\mathbf{C}$. Note that apart from the singular case $\sigma = 0$, $\Phi(\lambda, \mu)$ is found to be of order unity when $N \to \infty$. But because of the factor N in Eq. (19.6), the overlaps between $\mathbf{v}_i$ and $\mathbf{u}_j$ are of order $N^{-1/2}$ as soon as $\sigma > 0$.

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

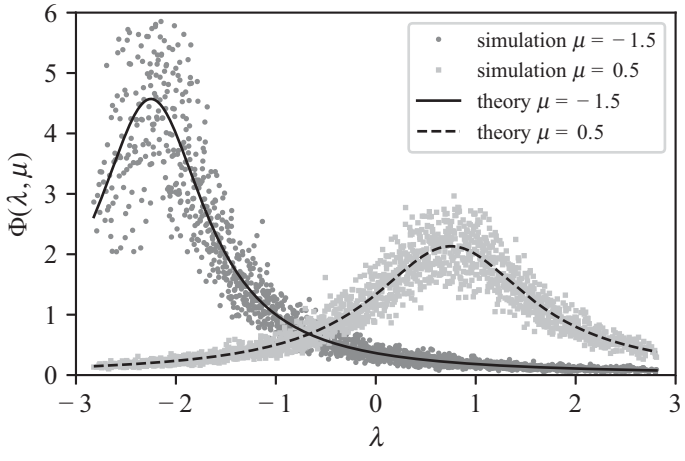

Figure 19.1 Normalized squared-overlap function $\Phi(\lambda, \mu)$ for $\mathbf{E} = \mathbf{X}_1 + \mathbf{X}_2$, the sum of two unit Wigner matrices, compared with a numerical simulation for $\mu = -1.5$ and $\mu = 0.5$. The simulations are for a single sample of size $N = 2000$, each data point corresponds to N times the square overlap between an eigenvector of $\mathbf{E}$ and one of $\mathbf{X}_1$ averaged over eigenvectors with eigenvalues within distance $\eta = 2/\sqrt{N}$ of μ.

Now suppose that $\mathbf{E} = \mathbf{C} + \mathbf{X}$ and $\mathbf{E}' = \mathbf{C} + \mathbf{X}'$, where $\mathbf{X}$ and $\mathbf{X}'$ are two independent Wigner matrices with the same variance σ^2. Using Eq. (19.8), one can compute the expected overlap between the eigenvectors of $\mathbf{E}$ and $\mathbf{E}'$. After a little work, one can establish the following result for $\Psi(\lambda, \lambda)$, i.e. the typical overlap around the same eigenvalues for $\mathbf{E}$ and $\mathbf{E}'$:

$$\Psi(\lambda, \lambda) = \frac{\sigma^2}{2 f_2(\lambda)^2} \frac{\partial_\lambda f_1(\lambda)}{(\partial_\lambda f_1(\lambda))^2 + (\partial_\lambda f_2(\lambda))^2}, \tag{19.14}$$

where

$$f_1(\lambda) = \lambda - \sigma^2 \mathfrak{h}_{\mathbf{E}}(\lambda); \qquad f_2(\lambda) = \sigma^2 \pi \rho_{\mathbf{E}}(\lambda); \qquad \mathfrak{h}_{\mathbf{E}}(\lambda) := \mathrm{Re}[\mathfrak{g}_{\mathbf{E}}(\lambda)]. \tag{19.15}$$

Note that in the large N limit, $\mathfrak{g}_{\mathbf{E}}(z) = \mathfrak{g}_{\mathbf{E}'}(z)$.

The formula for $\Psi(\lambda, \lambda')$ is more cumbersome; for a fixed λ', one finds again a humped shaped function with a maximum at $\lambda' \approx \lambda$. The most striking aspect of this formula, however, is that only $\mathfrak{g}_{\mathbf{E}}(z)$ (which is measurable from data) is needed to compute the expected overlap $\Psi(\lambda, \lambda')$; the knowledge of the "true" matrix $\mathbf{C}$ is not needed to judge whether or not the observed overlap between the eigenvectors of $\mathbf{E}$ and $\mathbf{E}'$ is compatible with the hypothesis that such matrices are both noisy versions of the same unknown $\mathbf{C}$.

19.1.3 Overlaps in the Multiplicative Case

We now repeat the same steps in the case where $\mathbf{E} = \mathbf{C}^{\frac{1}{2}} \mathbf{W}_q \mathbf{C}^{\frac{1}{2}}$, where $\mathbf{W}_q$ is a Wishart matrix of parameter q. We know that in this case the matrix subordination formula reads as (13.47), which can be rewritten as

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

$$\mathbf{G}_{\mathbf{E}}(z) = \frac{Z(z)}{z}\mathbf{G}_{\mathbf{C}}(Z(z)), \quad \text{with} \quad Z = \frac{z}{1 - q + qz\mathfrak{g}_{\mathbf{E}}(z)}. \tag{19.16}$$

This allows us to compute

$$\mathbf{u}_j^T \mathbf{G}_{\mathbf{E}}(\lambda_i - \mathrm{i}\eta)\mathbf{u}_j = \frac{Z(\lambda_i - \mathrm{i}\eta)}{\lambda_i - \mathrm{i}\eta} \frac{1}{Z(\lambda_i - \mathrm{i}\eta) - \mu_j}, \tag{19.17}$$

and finally, taking the imaginary part in the limit $\eta \to 0^+$,

$$\Phi(\lambda, \mu) = \frac{q\mu\lambda}{(\mu(1-q) - \lambda + q\mu\lambda\, \mathfrak{h}_{\mathbf{E}}(\lambda))^2 + q^2\mu^2\lambda^2\pi^2\rho_{\mathbf{E}}^2(\lambda)}, \tag{19.18}$$

where, again, $\mathfrak{h}_{\mathbf{E}}$ denotes the real part of the Stieltjes transform $\mathfrak{g}_{\mathbf{E}}$. Note that in the limit $q \to 0$, $\Phi(\lambda, \mu)$ becomes more and more peaked around $\lambda \approx \mu$, with an amplitude that diverges for $q = 0$. Indeed, in this limiting case, one should find that the sample eigenvectors $\mathbf{v}_i$ become equal to the population ones $\mathbf{u}_i$. More generally, $\Phi(\lambda, \mu)$ for a fixed μ has a Lorentzian humped shape as a function of λ, which peaks for $\lambda \approx \mu$.

Now suppose that $\mathbf{E} = \mathbf{C}^{\frac{1}{2}}\mathbf{W}_q\mathbf{C}^{\frac{1}{2}}$ and $\mathbf{E}' = \mathbf{C}^{\frac{1}{2}}\mathbf{W}'_q\mathbf{C}^{\frac{1}{2}}$, where $\mathbf{W}_q$ and $\mathbf{W}'_q$ are two independent Wishart matrices with the same parameter q. Using Eq. (19.8), one can again compute the expected overlap between the eigenvectors of $\mathbf{E}$ and $\mathbf{E}'$. The final formula is however too cumbersome to be reported here, see Bun et al. [2018]. The formula simplifies in the limit where $\mathbf{C}$ is close to the identity matrix, in the sense that $\tau(\mathbf{C}^2) = 1 + \epsilon$, with $\epsilon \to 0$. In this case:

$$\Psi(\lambda, \lambda') = 1 + \epsilon\,[2\,\mathfrak{h}_{\mathbf{E}}(\lambda) - 1]\left[2\,\mathfrak{h}_{\mathbf{E}}(\lambda') - 1\right] + O(\epsilon^2). \tag{19.19}$$

More generally, the squared overlaps only depend on $\mathfrak{g}_{\mathbf{E}}(z)$ (which is measurable from data). Again, the knowledge of the "true" matrix $\mathbf{C}$ is not needed to judge whether or not the observed overlap between the eigenvectors of $\mathbf{E}$ and $\mathbf{E}'$ is compatible with the hypothesis that such matrices are both noisy versions of the same unknown $\mathbf{C}$. This is particularly important in financial applications, where $\mathbf{E}$ and $\mathbf{E}'$ may correspond to covariance matrices measured on two non-overlapping periods. In such a case, the hypothesis that the true $\mathbf{C}$ is indeed the same in the two periods may not be warranted and can be directly tested using the overlap formula.

19.2 Rotationally Invariant Estimators

19.2.1 Setting the Stage

The results derived above concerning the overlaps between the eigenvectors of sample $\mathbf{E}$ and population (or "true") $\mathbf{C}$ matrices allow one to construct a rotationally invariant estimator of $\mathbf{C}$ knowing $\mathbf{E}$. The idea can be framed within the Bayesian approach of the previous chapter, when the prior knowledge about $\mathbf{C}$ is mute about the possible directions in which the eigenvectors of $\mathbf{C}$ are pointing. More formally, this can be expressed by saying that the prior distribution $P_0(\mathbf{C})$ is rotation invariant, i.e.

$$P_0(\mathbf{C}) = P_0(\mathbf{O}\mathbf{C}\mathbf{O}^T), \tag{19.20}$$

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

where $\mathbf{O}$ is an arbitray rotation matrix. Examples of rotationally invariant priors are provided by the orthogonal ensemble introduced in Chapter 5, where $P_0(\mathbf{C})$ only depends on $\mathrm{Tr}(\mathbf{C})$.

Now, since the posterior probability of $\mathbf{C}$ given $\mathbf{E}$ is given by

$$P(\mathbf{C}|\mathbf{E}) \propto (\det \mathbf{C})^{-T/2} \exp\left[-\frac{T}{2}\,\mathrm{Tr}\left(\mathbf{C}^{-1}\mathbf{E}\right)\right] P_0(\mathbf{C}), \tag{19.21}$$

it is easy to verify that the MMSE estimator of $\mathbf{C}$ transforms in the same way as $\mathbf{E}$ under an arbitrary rotation $\mathbf{O}$, i.e.

$$\begin{aligned}\mathbb{E}[\mathbf{C}|\mathbf{O}\mathbf{E}\mathbf{O}^T] &= \int \mathbf{C}\, P(\mathbf{C}|\mathbf{O}\mathbf{E}\mathbf{O}^T) P_0(\mathbf{C}) \mathcal{D}\mathbf{C} \\ &= \mathbf{O}\left[\int \widetilde{\mathbf{C}}\, P(\widetilde{\mathbf{C}}|\mathbf{E}) P_0(\widetilde{\mathbf{C}}) \mathcal{D}\widetilde{\mathbf{C}}\right]\mathbf{O}^T \\ &= \mathbf{O}\mathbb{E}(\mathbf{C}|\mathbf{E})\mathbf{O}^T, \end{aligned} \tag{19.22}$$

using the change of variable $\widetilde{\mathbf{C}} = \mathbf{O}^T\mathbf{C}\mathbf{O}$, and the explicit form of $P(\mathbf{C}|\mathbf{E})$ given in Eq. (19.21).

More generally, if we call $\Xi(\mathbf{E})$ an estimator of $\mathbf{C}$ given $\mathbf{E}$, this estimator is rotationally invariant if and only if

$$\Xi(\mathbf{O}\mathbf{E}\mathbf{O}^T) = \mathbf{O}\Xi(\mathbf{E})\mathbf{O}^T, \tag{19.23}$$

for any orthogonal matrix $\mathbf{O}$. This means in words that, if the SCM $\mathbf{E}$ is rotated by some $\mathbf{O}$, then our estimation of $\mathbf{C}$ must be rotated in the same fashion. Intuitively, this is because we have no prior assumption on the eigenvectors of $\mathbf{C}$, so the only special directions in which $\mathbf{C}$ can point are those singled out by $\mathbf{E}$ itself. Estimators abiding by Eq. (19.23) are called rotationally invariant estimators (RIE).

An alternative interpretation of Eq. (19.23) is that $\Xi(\mathbf{E})$ can be diagonalized in the same basis as $\mathbf{E}$, up to a *fixed* rotation matrix Ω. But consistent with our rotationally invariant prior on $\mathbf{C}$, there is no natural guess for Ω, except the identity matrix $\mathbf{1}$. Hence we conclude that $\Xi(\mathbf{E})$ has the same eigenvectors as those of $\mathbf{E}$, and write

$$\Xi(\mathbf{E}) = \sum_{i=1}^{N} \xi_i \mathbf{v}_i \mathbf{v}_i^T, \tag{19.24}$$

where $\mathbf{v}_i$ are, as above, the eigenvectors of $\mathbf{E}$, and where ξ_i is a function of all empirical eigenvalues $[\lambda_j]_{j\in(1,N)}$. We now show how these ξ_i can be optimally chosen, and operationally computed from data in the limit $N \to \infty$.

19.2.2 The Optimal RIE

Suppose we ask the following question: what is the optimal choice of ξ_i such that $\Xi(\mathbf{E})$, defined by Eq. (19.24), is as close as possible to the true $\mathbf{C}$? If the eigenvectors of $\mathbf{E}$ were

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

equal to those of $\mathbf{C}$, i.e. if $\mathbf{v}_i = \mathbf{u}_i$, $\forall i$, the solution would trivially be $\xi_i = \mu_i$. But in the case where $\mathbf{v}_i \neq \mathbf{u}_i$, the solution is *a priori* non-trivial. So we want to minimize the following least-square error:

$$\mathrm{Tr}(\Xi(\mathbf{E}) - \mathbf{C})^2 = \sum_{i=1}^{N} \mathbf{v}_i^T (\Xi(\mathbf{E}) - \mathbf{C})^2 \mathbf{v}_i = \sum_{i=1}^{N} \left(\xi_i^2 - 2\xi_i \mathbf{v}_i^T \mathbf{C} \mathbf{v}_i + \mathbf{v}_i^T \mathbf{C}^2 \mathbf{v}_i \right). \quad (19.25)$$

Minimizing over ξ_k and noting that the third term in the equation above is independent of the ξ's, it is easy to get the following expression for the optimal ξ_k:

$$\xi_k = \mathbf{v}_k^T \mathbf{C} \mathbf{v}_k. \quad (19.26)$$

This is all very well but seems completely absurd: we assume that we do not know the true $\mathbf{C}$ and want to find the best estimator of $\mathbf{C}$ knowing $\mathbf{E}$, and we find an equation for the ξ that we cannot compute unless we know $\mathbf{C}$.

Because Eq. (19.26) requires in principle knowledge we do not have, it is often called the "oracle" estimator. But as we will see in the next section, the large N limit allows one to actually compute the optimal ξ's from the data alone, without having to know $\mathbf{C}$.

19.2.3 The Large Dimension Miracle

Let us first rewrite Eq. (19.26) in terms of the overlaps introduced in Section 19.1. Expanding over the eigenvectors of $\mathbf{C}$ we find

$$\xi_k = \sum_{j=1}^{N} \mathbf{v}_k^T \mathbf{u}_j \mu_j \mathbf{u}_j^T \mathbf{v}_k = \sum_{j=1}^{N} \mu_j \left(\mathbf{u}_j^T \mathbf{v}_k \right)^2 \quad (19.27)$$

$$\xrightarrow[N\to\infty]{} \int \mathrm{d}\mu \, \rho_{\mathbf{C}}(\mu) \mu \Phi(\lambda_k, \mu), \quad (19.28)$$

where λ_k is the eigenvalue of the sample matrix $\mathbf{E}$ associated with $\mathbf{v}_k$. In other words, ξ_k is an average over the eigenvalues of $\mathbf{C}$, weighted by the square overlaps $\Phi(\lambda_k, \mu)$. Now, using Eq. (19.5), we can also write

$$\begin{aligned} \xi_k = \sum_{j=1}^{N} \mu_j \left(\mathbf{u}_j^T \mathbf{v}_k \right)^2 &= \frac{1}{\pi \rho_{\mathbf{E}}(\lambda_k)} \lim_{\eta \to 0^+} \mathrm{Im} \sum_{j=1}^{N} \mathbf{u}_j^T \mu_j \mathbf{G}_{\mathbf{E}}(\lambda_k - \mathrm{i}\eta) \mathbf{u}_j \\ &= \frac{1}{\pi \rho_{\mathbf{E}}(\lambda_k)} \lim_{\eta \to 0^+} \mathrm{Im}\, \tau \left(\mathbf{C} \mathbf{G}_{\mathbf{E}}(\lambda_k - \mathrm{i}\eta) \right). \end{aligned} \quad (19.29)$$

We now use the fact that in both the additive and the multiplicative case, the matrices $\mathbf{G}_{\mathbf{E}}$ and $\mathbf{G}_{\mathbf{C}}$ are related by a subordination equation of the type

$$\mathbf{G}_{\mathbf{E}}(z) = Y(z) \mathbf{G}_{\mathbf{C}}(Z(z)), \quad (19.30)$$

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

with $Y(z) = 1$ in the additive case and $Y(z) = Z(z)/z$ in the multiplicative case. Hence we can write the following series of equalities:

$$\begin{aligned}\tau\left(\mathbf{C}\mathbf{G}_{\mathbf{E}}(z)\right) &= Y\tau\left(\mathbf{C}\mathbf{G}_{\mathbf{C}}(Z)\right) = Y\tau\left((\mathbf{C} - Z\mathbf{1} + Z\mathbf{1})(Z\mathbf{1} - \mathbf{C})^{-1}\right) \\ &= YZ\mathfrak{g}_{\mathbf{C}}(Z) - Y = Z(z)\mathfrak{g}_{\mathbf{E}}(z) - Y(z). \end{aligned} \tag{19.31}$$

But since $Z(z)$ only depend on $\mathfrak{g}_{\mathbf{E}}(z)$, we see that the final formula for ξ_k does not explicitly depend on $\mathbf{C}$ anymore and reads

$$\xi_k = \frac{1}{\pi\rho_{\mathbf{E}}(\lambda_k)} \lim_{\eta\to 0^+} Z(z_k)\mathfrak{g}_{\mathbf{E}}(z_k) - Y(z_k), \qquad z_k := \lambda_k - \mathrm{i}\eta. \tag{19.32}$$

Since all the quantities on the right hand side can be estimated from the data alone, this formula will lend itself to real world applications. Let us first explore this formula for two simple cases, for an additive model and for a multiplicative model.

19.2.4 The Additive Case

For a general noise matrix $\mathbf{X}$, one has $Z(z) = z - R_{\mathbf{X}}(\mathfrak{g}_{\mathbf{E}}(z))$, leading to the following mapping between the empirical eigenvalues λ and the RIE eigenvalues ξ:

$$\xi(\lambda) = \lambda - \frac{\lim_{\eta\to 0^+} \mathrm{Im}\, R_{\mathbf{X}}(\mathfrak{g}_{\mathbf{E}}(z))\mathfrak{g}_{\mathbf{E}}(z)}{\lim_{\eta\to 0^+} \mathrm{Im}\, \mathfrak{g}_{\mathbf{E}}(z)}, \quad z = \lambda - \mathrm{i}\eta. \tag{19.33}$$

If there is no noise, i.e. $\mathbf{X} = 0$ and hence $R_{\mathbf{X}} = 0$, we find as expected $\xi(\lambda) = \lambda$. If $\mathbf{X}$ is small, then

$$R_{\mathbf{X}}(x) = \epsilon x + \cdots, \tag{19.34}$$

where we have assumed $\tau(\mathbf{X}) = 0$ and $\epsilon = \tau(\mathbf{X}^2)$ is small. Hence we find

$$\xi(\lambda) = \lambda - 2\epsilon\, \mathfrak{h}_{\mathbf{E}}(\lambda) + \cdots. \tag{19.35}$$

A natural case to consider is when $\mathbf{X}$ is Wigner noise, for which $R_{\mathbf{X}}(x) = \sigma_{\mathrm{n}}^2 x$ exactly, such that the equation above is exact with $\epsilon = \sigma_{\mathrm{n}}^2$, for arbitrary values of σ_{n}. When $\mathbf{C}$ is another Wigner matrix with variance σ_{s}^2, then $\mathbf{E}$ is clearly also a Wigner matrix with variance $\sigma^2 = \sigma_{\mathrm{n}}^2 + \sigma_{\mathrm{s}}^2$. In this case, when $-2\sigma < \lambda < 2\sigma$,

$$\mathfrak{h}_{\mathbf{E}}(\lambda) = \frac{\lambda}{2\sigma^2}. \tag{19.36}$$

Hence we obtain, from Eq. (2.38),

$$\xi(\lambda) = \lambda - \frac{\lambda\sigma_{\mathrm{n}}^2}{\sigma^2} = r\lambda, \qquad r := \frac{\sigma_{\mathrm{s}}^2}{\sigma^2}, \tag{19.37}$$

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

which is the linear shrinkage obtained for Gaussian variables in Chapter 18. In fact, this shrinkage formula is expected elementwise, since all elements are Gaussian random variables:[1]

$$\Xi(\mathbf{E})_{ij} = r\,\mathbf{E}_{ij}, \tag{19.38}$$

see Eq. (18.10) with $x_0 = 0$.

Exercise 19.2.1 Additive RIE for the sum of two matrices from the same distribution

In this exercise we will find a simple form for a RIE estimator when the noise is drawn from the same distribution as the signal, i.e.

$$\mathbf{E} = \mathbf{C} + \mathbf{X}, \tag{19.39}$$

with $\mathbf{X}$ and $\mathbf{C}$ mutually free matrices drawn from the same ensemble.

(a) Write a relationship between $R_{\mathbf{X}}(g)$ and $R_{\mathbf{E}}(g)$.
(b) Given that $\mathfrak{g}_{\mathbf{E}}(z)R_{\mathbf{E}}(\mathfrak{g}_{\mathbf{E}}(z)) = z\mathfrak{g}_{\mathbf{E}}(z) - 1$, what is $\mathfrak{g}_{\mathbf{E}}(z)R_{\mathbf{X}}(\mathfrak{g}_{\mathbf{E}}(z))$?
(c) Use Eq. (19.33) and the fact that z is real in the limit $\eta \to 0^+$ to show that $\xi(\lambda) = \lambda/2$.
(d) Given that $\Xi = \mathbb{E}[\mathbf{C}]_{\mathbf{E}}$ (see Section 19.4), find a simple symmetry argument to show that $\Xi = \mathbf{E}/2$.
(e) Generate numerically two independent symmetric orthogonal matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ with $N = 1000$ (see Exercise 1.2.4). Compute the eigenvalues λ_k and eigenvectors $\mathbf{v}_k$ of the sum of these two matrices.
(f) Plot the normalized histogram of the λ_k's and compare with the arcsine law between -2 and 2 ($\rho(\lambda) = 1/(\pi\sqrt{4-\lambda^2})$).
(g) Make a scatter plot $\mathbf{v}_k^T\mathbf{M}_1\mathbf{v}_k$ vs λ_k and compare with $\lambda/2$.

19.2.5 The Multiplicative Case

We can now tackle the multiplicative case, which includes the important practical problem of estimating the true covariance matrix given a sample covariance matrix. In the multiplicative case, it is more elegant to use the subordination relation Eq. (13.46) for the T-matrix rather than for the resolvent. In the present setting we thus write

$$\mathbf{T}_{\mathbf{E}}(z) = \mathbf{T}_{\mathbf{C}}[z\,S_{\mathbf{W}}(\mathfrak{t}_{\mathbf{E}}(z))]. \tag{19.40}$$

[1] Note however that there is a slight subtlety here: the linear shrinkage equation (19.37) only holds in the absence of outliers, i.e. empirical eigenvalues that fall outside the interval $(-2\sigma, 2\sigma)$. For such eigenvalues, shrinkage is non-linear. For a similar situation in the multiplicative case, see Figure 19.2.

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

In terms of T-transforms, Eq. (19.29) reads

$$\xi(\lambda) = \frac{\lim_{\eta\to 0^+} \operatorname{Im} \tau\left(\mathbf{C}\mathrm{t}_{\mathbf{C}}[z\, S_{\mathbf{W}}(\mathrm{t}_{\mathbf{E}}(z))]\right)}{\lim_{\eta\to 0^+} \operatorname{Im} \mathrm{t}_{\mathbf{E}}(z)}; \qquad z = \lambda - \mathrm{i}\eta. \tag{19.41}$$

Since $\mathbf{T}_{\mathbf{C}}(z) = \mathbf{C}(z\mathbf{1} - \mathbf{C})^{-1}$, we have, with $t = \mathrm{t}_{\mathbf{E}}(z)$ as a shorthand,

$$\begin{aligned}
\tau\left[\mathbf{C}\mathrm{t}_{\mathbf{C}}(zS_{\mathbf{W}}(t))\right] &= \tau\left[\mathbf{C}^2(zS_{\mathbf{W}}(t)\mathbf{1} - \mathbf{C})^{-1}\right] \\
&= \tau\left[\mathbf{C}(\mathbf{C} - zS_{\mathbf{W}}(t)\mathbf{1} + zS_{\mathbf{W}}(t)\mathbf{1})(zS_{\mathbf{W}}(t)\mathbf{1} - \mathbf{C})^{-1}\right] \\
&= \tau(\mathbf{C}) + zS_{\mathbf{W}}(t)\mathrm{t}_{\mathbf{C}}(zS_{\mathbf{W}}(t)) \\
&= \tau(\mathbf{C}) + zS_{\mathbf{W}}(t)\mathrm{t}_{\mathbf{E}}(z).
\end{aligned} \tag{19.42}$$

The first term $\tau(\mathbf{C})$ is real and does not contribute to the imaginary part that we have to compute, so we obtain

$$\xi(\lambda) = \lambda \frac{\lim_{\eta\to 0^+} \operatorname{Im} S_{\mathbf{W}}(\mathrm{t}_{\mathbf{E}}(z))\mathrm{t}_{\mathbf{E}}(z)}{\lim_{\eta\to 0^+} \operatorname{Im} \mathrm{t}_{\mathbf{E}}(z)}, \quad z = \lambda - \mathrm{i}\eta. \tag{19.43}$$

Equation (19.43) is very general. It applies to sample covariance matrices where the noise matrix $\mathbf{W}$ is a white Wishart, but it also applies to more general multiplicative noise processes.

In the special case of sample covariance matrices $\mathbf{E} = \mathbf{C}^{\frac{1}{2}}\mathbf{W}_q\mathbf{C}^{\frac{1}{2}}$ with $N/T = q$, we know that $S_{\mathbf{W}_q}(t) = (1+qt)^{-1}$. In the bulk region $\lambda_- < \lambda < \lambda_+$, $t = \mathrm{t}_{\mathbf{E}}(z)$ is complex with non-zero imaginary part when $z = \lambda - \mathrm{i}\eta$. Hence

$$\xi(\lambda) = \lambda \frac{\lim_{\eta\to 0^+} \operatorname{Im} \frac{t}{1+qt}}{\lim_{\eta\to 0^+} \operatorname{Im} t} = \left.\frac{\lambda}{|1 + q\mathrm{t}_{\mathbf{E}}(\lambda - \mathrm{i}\eta)|^2}\right|_{\eta\to 0}, \tag{19.44}$$

where we have used the fact that

$$\operatorname{Im} \frac{t}{1+qt} = \operatorname{Im} \frac{t(1+qt^\star)}{|1+qt|^2} = \operatorname{Im} \frac{t + q|t|^2}{|1+qt|^2} = \frac{1}{|1+qt|^2} \operatorname{Im} t. \tag{19.45}$$

Equation (19.44) can be interpreted as a form of non-linear shrinkage. A way to see this is to note that below λ_- and above λ_+ (the edges of the sample spectrum) $\mathrm{t}_{\mathbf{E}}(\lambda)$ is real. From the very definition,

$$\mathrm{t}_{\mathbf{E}}(\lambda) = \int_{\lambda_-}^{\lambda_+} \mathrm{d}\lambda' \rho_{\mathbf{E}}(\lambda') \frac{\lambda'}{\lambda - \lambda'} = \lambda \mathrm{g}_{\mathbf{E}}(\lambda) - 1, \tag{19.46}$$

for any λ outside or at the edges of the spectrum. Hence, since $\lambda_- \geq 0$ for covariance matrices, $\mathrm{t}_{\mathbf{E}}(\lambda_-) < 0$ and $\mathrm{t}_{\mathbf{E}}(\lambda_+) > 0$. Hence, one directly establishes that the support of the RIE $\Xi(\mathbf{E})$ is narrower than that of $\mathbf{E}$:

$$\xi(\lambda_-) \geq \lambda_-; \qquad \xi(\lambda_+) \leq \lambda_+, \tag{19.47}$$

where the inequalities are saturated for $q = 0$, in which case, as expected $\xi(\lambda) = \lambda$, $\forall \lambda \in (\lambda_-, \lambda_+)$. A more in-depth discussion of the properties of Eq. (19.44) is given in Section 19.3.

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

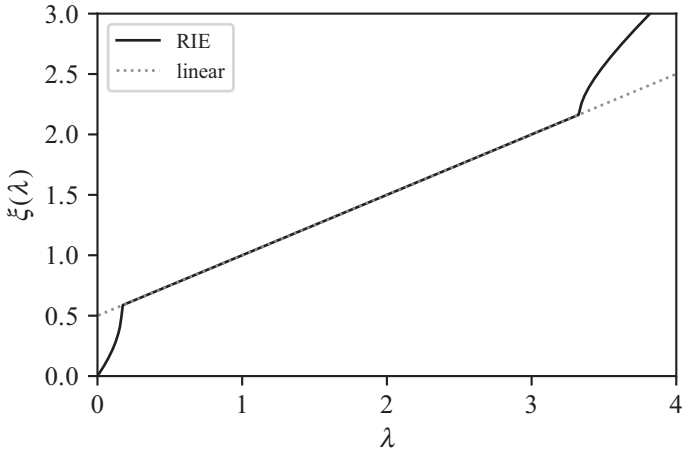

Figure 19.2 The RIE estimator (19.44) for a true covariance matrix given by an inverse-Wishart of variance $p = 0.25$ observed using data with aspect ratio $q = 0.25$. On the support of the sample density ($\lambda \in [0.17, 3.33]$), the RIE matches perfectly the linear shrinkage estimator (19.49) (with $r = 1/2$), but it is different from it outside of the expected spectrum.

Using Eq. (19.46), the shrinkage equation (19.44) can be rewritten as

$$\xi(\lambda) = \left. \frac{\lambda}{|1 - q + q\lambda g_{\mathbf{E}}(\lambda - i\eta)|^2} \right|_{\eta \to 0^+}, \tag{19.48}$$

a result first derived in Ledoit and Péché [2011].

Equation (19.44) considerably simplifies in the case where the true covariance matrix $\mathbf{C}$ is an inverse-Wishart matrix of parameter p. Injecting the explicit form of $t_{\mathbf{E}}(z)$ given by Eq. (15.50) into Eq. (19.44) leads, after simple manipulations, to

$$\xi(\lambda) = \frac{q + \lambda p}{p + q} = r\lambda + (1 - r); \qquad r := \frac{p}{p + q}, \tag{19.49}$$

i.e. exactly the linear shrinkage result derived in a Bayesian framework in Section 18.3. Note that the result Eq. (19.49) only holds between λ_- and λ_+, given in Eq. (15.52). The full function $\xi(\lambda)$ when $\mathbf{C}$ is an inverse-Wishart matrix is given in Figure 19.2.

Exercise 19.2.2 RIE when the true covariance matrix is Wishart

Assume that the true covariance matrix $\mathbf{C}$ is given by a Wishart matrix with parameter q_0. This case is a tractable model for which the computation can be done semi-analytically (we will get cubic equations!).

We observe a sample covariance matrix $\mathbf{E}$ over $T = qN$ time intervals. $\mathbf{E}$ is the free product of $\mathbf{C}$ and another Wishart matrix of parameter q:

$$\mathbf{E} = \mathbf{C}^{\frac{1}{2}} \mathbf{W} \mathbf{C}^{\frac{1}{2}}. \tag{19.50}$$

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

(a) Given that the S-transform of the true covariance is $S_{\mathbf{C}}(t) = 1/(1+q_0 t)$ and the S-transform of the Wishart is $S_{\mathbf{W}}(t) = 1/(1+qt)$, use the product of S-transforms for the free product and Eq. (11.92) to write an equation for $t_{\mathbf{E}}(z)$. It should be a cubic equation in t.

(b) Using a numerical polynomial solver (e.g. np.roots) solve for $\mathrm{t}_{\mathbf{E}}(z)$ for z real between 0 and 4, choose $q_0 = 1/4$ and $q = 1/2$. Choose the root with positive imaginary part. Use Eqs. (11.89) and (2.47) to find the eigenvalue density and plot this density. The edge of the spectrum should be (slightly below) 0.05594 and (slightly above) 3.746.

(c) For λ in the range [0.05594, 3.746] plot the optimal cleaning function (use the same solution $\mathrm{t}_{\mathbf{E}}(z)$ as in (b)):

$$\xi(\lambda) = \frac{\lambda}{|1+qt(\lambda)|^2}. \tag{19.51}$$

(d) For $N = 1000$ numerically generate $\mathbf{C}$ ($q_0 = 1/4$), two versions of $\mathbf{W}_{1,2}$ ($q = 1/2$) and hence two versions of $\mathbf{E}_{1,2} := \mathbf{C}^{\frac{1}{2}}\mathbf{W}_{1,2}\mathbf{C}^{\frac{1}{2}}$. $\mathbf{E}_1$ will be the "in-sample" matrix and $\mathbf{E}_2$ the "out-of-sample" matrix. Check that $\tau(\mathbf{C}) = \tau(\mathbf{W}_{1,2}) = \tau(\mathbf{E}_{1,2}) = 1$ and that $\tau(\mathbf{C}^2) = 1.25$, $\tau(\mathbf{W}_{1,2}^2) = 1.5$ and $\tau(\mathbf{E}_{1,2}^2) = 1.75$.

(e) Plot the normalized histogram of the eigenvalues of $\mathbf{E}_1$, it should match your plot in (b).

(f) For every eigenvalue, eigenvector pair $(\lambda_k, \mathbf{v}_k)$ of $\mathbf{E}_1$ compute $\xi_{\mathrm{val}}(\lambda_k) := \mathbf{v}_k^T \mathbf{E}_2 \mathbf{v}_k$. Plot $\xi_{\mathrm{val}}(\lambda_k)$ vs λ_k and compare with your answer in (c).

Exercise 19.2.3 Multiplicative RIE when the signal and the noise have the same distribution

(a) Adapt the arguments of Exercise 19.2.1 to the multiplicative case with $\mathbf{C}$ and $\mathbf{W}$ two free matrices drawn from the same ensemble. Show that in this case $\xi(\lambda) = \sqrt{\lambda}$.

(b) Redo Exercise 19.2.2 with $q = q_0 = 1/4$, compare your $\xi(\lambda)$ with $\sqrt{\lambda}$.

19.2.6 RIE for Outliers

So far, we have focused on "cleaning" the bulk eigenvectors. But it turns out that the formulas above are also valid for outliers of $\mathbf{C}$ that appear as outliers of $\mathbf{E}$. One can show that, outside the bulk, $\mathrm{g}_{\mathbf{E}}(z)$ and $\mathrm{t}_{\mathbf{E}}(z)$ are analytic on the real axis and thus, for small η,

$$\mathrm{Im}\, \mathrm{g}_{\mathbf{E}}(\lambda - \mathrm{i}\eta) = -\eta \mathrm{g}'_{\mathbf{E}}(\lambda), \qquad \mathrm{Im}\, \mathrm{t}_{\mathbf{E}}(\lambda - \mathrm{i}\eta) = -\eta \mathrm{t}'_{\mathbf{E}}(\lambda). \tag{19.52}$$

Then Eqs. (19.33) and (19.43) simplify to

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

$$\begin{aligned} \xi(\lambda) &= \lambda - \frac{\mathrm{d}}{\mathrm{d}g}\left[gR_{\mathbf{X}}(g)\right], \qquad g = \mathfrak{g}_{\mathbf{E}}(\lambda), \\ \xi(\lambda) &= \lambda \frac{\mathrm{d}}{\mathrm{d}t}\left[tS_{\mathbf{W}}(t)\right], \qquad t = \mathfrak{t}_{\mathbf{E}}(\lambda), \end{aligned} \tag{19.53}$$

respectively for the additive and multiplicative cases.

19.3 Properties of the Optimal RIE for Covariance Matrices

Even though the optimal non-linear shrinkage function (19.44), (19.48) seems relatively simple, it is not immediately clear what is the effect induced by the transformation $\lambda_i \to \xi(\lambda_i)$. In this section, we thus give some quantitative properties of the optimal estimator Ξ to understand the impact of the optimal non-linear shrinkage function.

First let us consider the moments of the spectrum of Ξ. From Eqs. (19.24) and (19.26) we immediately derive that

$$\mathrm{Tr}\,\Xi = \sum_{j=1} \mu_j \mathbf{u}_j^T \left(\sum_{i=1} \mathbf{v}_i \mathbf{v}_i^T \right) \mathbf{u}_j = \mathrm{Tr}\,\mathbf{C}, \tag{19.54}$$

meaning that the cleaning operation preserves the trace of the population matrix $\mathbf{C}$, as it should do. For the second moment, we have

$$\mathrm{Tr}\,\Xi^2 = \sum_{j,k=1}^{N} \mu_j \mu_k \sum_{i=1}^{N} (\mathbf{v}_i^T \mathbf{u}_j)^2 (\mathbf{v}_i^T \mathbf{u}_k)^2.$$

Now, if we define the matrix $\mathbf{A}_{jk}$ as $\sum_{i=1}^{N} (\mathbf{v}_i^T \mathbf{u}_j)^2 (\mathbf{v}_i^T \mathbf{u}_k)^2$ for $j,k = 1,N$, it is not hard to see that it is a matrix with non-negative entries and whose rows all sum to unity (remember that all $\mathbf{v}_i$'s are normalized to unity). The matrix $\mathbf{A}$ is therefore a (bi-)stochastic matrix and the Perron–Frobenius theorem tells us that its largest eigenvalue is equal to unity (see Section 1.2.2). Hence, we deduce the following general inequality:

$$\sum_{j,k=1}^{N} \mathbf{A}_{j,k}\, \mu_j \mu_k \leq \sum_{j=1}^{N} \mu_j^2,$$

which implies that

$$\mathrm{Tr}\,\Xi^2 \leq \mathrm{Tr}\,\mathbf{C}^2 \leq \mathrm{Tr}\,\mathbf{E}^2, \tag{19.55}$$

where the last inequality comes from Eq. (17.11). In words, this result states that the spectrum of Ξ is narrower than the spectrum of $\mathbf{C}$, which is itself narrower than the spectrum of $\mathbf{E}$. The optimal RIE therefore tells us that we had better be even more cautious than simply bringing back the sample eigenvalues to their estimated true locations. This is because we have only partial information about the true eigenbasis of $\mathbf{C}$. In particular, one should always shrink downward (resp. upward) the small (resp. top) eigenvalues compared to their true locations μ_i for any $i \in (1,N)$, except for the trivial case $\mathbf{C} = \mathbf{1}$.

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

Next, we consider the asymptotic behavior of the optimal non-linear shrinkage function (19.44), (19.48). Throughout the following, suppose that we have an outlier at the left of the lower bound of $\rho_{\mathbf{E}}$ and let us assume $q < 1$ so that $\mathbf{E}$ has no exact zero mode. We know from Section 19.2.6 that the estimator (19.44) holds for outliers. Moreover, we have that $\lambda \mathrm{g}_{\mathbf{E}}(\lambda) = O(\lambda)$ for $\lambda \to 0$. This allows us to conclude from Eq. (19.26) that, for outliers very close to zero,

$$\xi(\lambda) = \frac{\lambda}{(1-q)^2} + O(\lambda^2), \tag{19.56}$$

which is in agreement with Eq. (19.55): small eigenvalues must be pushed upwards for $q > 0$.

The other asymptotic limit $\lambda \to \infty$ is also useful since it gives us the behavior of the non-linear shrinkage function ξ for large outliers. In that case, we know from Eq. (17.8) that $\lim_{\lambda\to\infty} \lambda \mathrm{t}_{\mathbf{E}}(\lambda) \sim \lambda^{-1}\tau(\mathbf{E})$. Therefore, we conclude that

$$\xi(\lambda) \approx \frac{\lambda}{\Big(1 + q\lambda^{-1}\tau(\mathbf{E}) + O(\lambda^{-2})\Big)^2} \approx \lambda - 2q\tau(\mathbf{E}) + O(\lambda^{-1}). \tag{19.57}$$

If all variances are normalized to unity such that $\tau(\mathbf{E}) = \tau(\mathbf{C}) = 1$, then we simply obtain

$$\xi(\lambda) \approx \lambda - 2q + O(\lambda^{-1}). \tag{19.58}$$

It is interesting to compare this with Eq. (14.54) for large rank-1 perturbations, which gives $\lambda \approx \mu + q$ for $\lambda \to \infty$. As a result, we deduce from Eq. (19.58) that $\xi(\lambda) \approx \mu - q$ and we therefore find the following ordering relation:

$$\xi(\lambda) < \mu < \lambda, \tag{19.59}$$

for isolated and large eigenvalues λ and for $q > 0$. Again, this result is in agreement with Eq. (19.55): large eigenvalues should be reduced downward for any $q > 0$, even below the "true" value of the outlier μ. More generally, the non-linear shrinkage function ξ interpolates smoothly between $\lambda/(1-q)^2$ for small λ to $\lambda - 2q$ for large λ.

19.4 Conditional Average in Free Probability

In this section we give an alternative derivation of the RIE formula, Eq. (19.29). This derivation is more elegant, albeit more abstract. In particular, it does not rely on the computation of eigenvector overlap, so by itself it misses the important link between the RIE and the computation of overlaps.

In the context of free probability, we work with abstract objects ($\mathbf{E}$, $\mathbf{C}$, etc.) that satisfy the axioms of Chapter 11. We can think of them as infinite-dimensional matrices. We are given the matrix $\mathbf{E}$ that was obtained by free operations from an unknown matrix $\mathbf{C}$. For instance it could be given by a combination of free product and free sum.

The matrix $\mathbf{E}$ is generated from the matrix $\mathbf{C}$; in this sense, $\mathbf{E}$ depends on $\mathbf{C}$. We would like to find the best estimator (in the least-square sense) of $\mathbf{C}$ given $\mathbf{E}$. It is given by the conditional average

$$\Xi = \mathbb{E}[\mathbf{C}]_{\mathbf{E}}. \tag{19.60}$$

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

In this abstract context, the only object we know is $\mathbf{E}$ so Ξ must be a function of $\mathbf{E}$. Let us call this function $\Xi(\mathbf{E})$. The fact that Ξ is a function of $\mathbf{E}$ only imposes that Ξ commutes with $\mathbf{E}$, i.e. that Ξ is diagonal in the eigenbasis of $\mathbf{E}$. One way to determine the function $\Xi(\mathbf{E})$ is to compute all possible moments of the form $m_k = \tau[\Xi(\mathbf{E})\mathbf{E}^k]$. They can be combined in the function

$$F(z) := \tau\left[\xi(\mathbf{E})(z\mathbf{1} - \mathbf{E})^{-1}\right] \tag{19.61}$$

via its Taylor series at $z \to \infty$. Using Eq. (19.60), we write

$$F(z) = \tau\left[\mathbb{E}[\mathbf{C}]|_{\mathbf{E}}\,(z\mathbf{1} - \mathbf{E})^{-1}\right]. \tag{19.62}$$

But the operator $\tau[.]$ contains the expectation value over all variables, both trace and randomness. So by the law of total expectation, $\tau(\mathbb{E}[.]) = \tau(.)$ and

$$F(z) = \tau\left[\mathbf{C}(z\mathbf{1} - \mathbf{E})^{-1}\right]. \tag{19.63}$$

To recover the function $\xi(\lambda)$ from $F(z)$ we use a spectral decomposition of $\mathbf{E}$:

$$F(z) = \int \rho_{\mathbf{E}}(\lambda)\frac{\xi(\lambda)}{z - \lambda}\mathrm{d}\lambda, \tag{19.64}$$

so

$$\lim_{\eta \to 0^+} \operatorname{Im} F(\lambda - \mathrm{i}\eta) = \pi\rho_{\mathbf{E}}(\lambda)\xi(\lambda), \tag{19.65}$$

which is equivalent to

$$\xi(\lambda) = \lim_{\eta \to 0^+} \frac{\operatorname{Im} F(\lambda - \mathrm{i}\eta)}{\operatorname{Im} \mathrm{g}_{\mathbf{E}}(\lambda - \mathrm{i}\eta)}, \tag{19.66}$$

itself equivalent to Eq. (19.29).

19.5 Real Data

As stated above, the good news about the RIE estimator is that it only depends on transforms of the observable matrix $\mathbf{E}$, such as $\mathrm{g}_{\mathbf{E}}(z)$ and $\mathrm{t}_{\mathbf{E}}(z)$ and the R- or S-transform of the noise process. One may think that real world applications should be relatively straightforward. However, we need to know the behavior of the *limiting* transforms on the real axis, precisely where the discrete N transforms $g_N(z)$ and $\mathrm{t}_N(z)$ fail to converge.

We will discuss here how to compute these transforms using either a parametric fit or a non-parametric approximation on the sample eigenvalues. In both cases we will tackle the multiplicative case with a Wishart noise but the discussion can be adapted to cover the additive case or any other type of noise. In Section 19.6 we will discuss an alternative approach using two datasets (or disjoint subsets of the original data).

19.5.1 Parametric Approach

Ansatz on $\rho_{\mathbf{C}}$ or $S_{\mathbf{C}}$

One can postulate a convenient functional form for $\rho_{\mathbf{C}}(\lambda)$ and fit the associated parameters on the data. This allows one to obtain analytical formulas for all the relevant transforms, from which one can extract the exact behavior on the real axis.

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

The simplest (most tractable) choice for $\rho_{\mathbf{C}}(\lambda)$ is the inverse-Wishart distribution. In this case $\rho_{\mathbf{E}}(\lambda)$ can be computed exactly (see Eq. (15.51)) and the optimal estimator is linear within the bulk of the spectrum, cf. Eq. (19.49). When the sample covariance matrix is normalized such that $\tau(\mathbf{E}) = 1$, the inverse-Wishart has a single parameter p that needs to be estimated from the data. As an estimate, one can use for example the second moment of $\mathbf{E}$:

$$\tau\left(\mathbf{E}^2\right) = 1 + p + q, \tag{19.67}$$

or its first inverse moment:

$$\tau\left(\mathbf{E}^{-1}\right) = \frac{1+p}{1-q}, \tag{19.68}$$

which is obtained using Eq. (15.13) with $S_{\mathbf{E}}(t) = (1 - pt)/(1 + qt)$, or simply by noting that $\tau(\mathbf{W}_q^{-1}\mathbf{M}_p^{-1}) = \tau(\mathbf{W}_q^{-1})\tau(\mathbf{M}_p^{-1})$ for free matrices, and using the results of Sections 15.2.2 and 15.2.3.

When the distribution of sample eigenvalues appears to be bounded from above and below, one can use a more complicated but still relatively tractable ansatz for $\rho_{\mathbf{C}}(\lambda)$, by postulating a simple form for its S-transform. For example using

$$S_{\mathbf{C}}(t) = \frac{(1-p_1 t)(1-p_2 t)}{1+q_1 t} \quad \Leftrightarrow \quad S_{\mathbf{E}}(t) = \frac{(1-p_1 t)(1-p_2 t)}{(1+qt)(1+q_1 t)}, \tag{19.69}$$

one finds that $\mathfrak{t}_{\mathbf{E}}(\zeta)$ (and hence $\rho_{\mathbf{E}}(\lambda)$) is the solution of a cubic equation. Higher order terms in t in the numerator or the denominator will give higher order equations for $\mathfrak{t}_{\mathbf{E}}(\zeta)$. The parameters p_1, p_2, q_1, etc. can be evaluated from the first few moments and inverse moments of $\mathbf{E}$ or by fitting the observed density of eigenvalues. However, the particularly convenient choice Eq. (19.69) does not work when the observed distribution of eigenvalues does not have enough skewness, as in the example shown in Figure 19.3.

Parametric Fit of $\rho_{\mathbf{E}}$

Another approach consists of postulating a form for the density of sample eigenvalues and fitting its parameters. For example, one can postulate that

$$\rho_{\mathbf{E}}(\lambda) = Z^{-1}\frac{(1+a_1\lambda+a_2\lambda^2)\sqrt{(\lambda-\lambda_-)(\lambda_+-\lambda)}}{1+b_1\lambda+b_2\lambda^2}, \tag{19.70}$$

where $\lambda_\pm$ are fixed to the smallest/largest observed eigenvalues, and a_1, a_2, b_1 and b_2 are fitted on the data by minimizing the square error on the cumulative distribution. The normalization factor Z can be computed during the fitting procedure. This particular form fits very well for sample data generated numerically (see Fig. 19.3 left). To find the optimal shrinkage function (19.48), we then reconstruct the complex Stieltjes transform $\mathfrak{g}_{\mathbf{E}}(x - \mathrm{i}0^+)$ numerically, by using the fitted $\rho(\lambda)$ and computing its Hilbert transform. The issue with such an approach is that even when Eq. (19.70) is a good fit to the sample density of eigenvalues, it cannot be obtained as the result of the free product of a Wishart and some

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

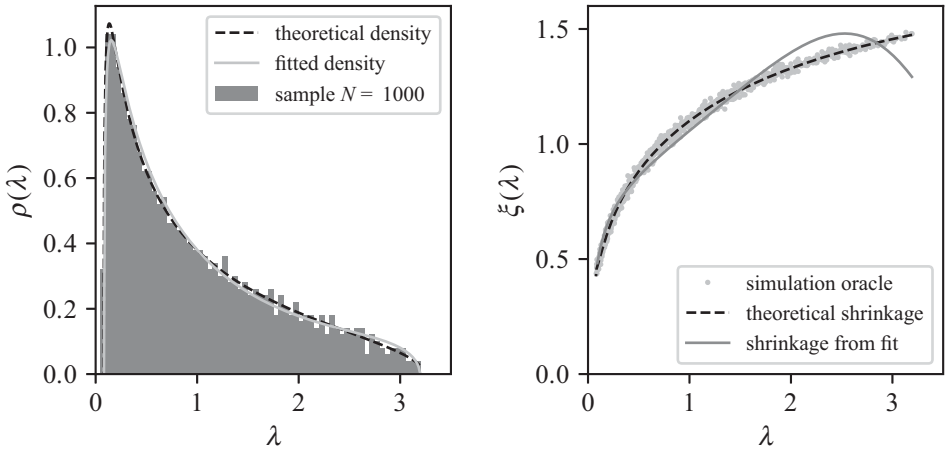

Figure 19.3 Parametric fit illustrated on an example where the true covariance has a uniform density of eigenvalues with mean 1 and variance 0.2 (see Eq. (15.42)). A single sample covariance matrix with $N = 1000$ and $q = 0.4$ was generated, and the ad-hoc distribution (19.70) was fitted to the eigenvalue CDF. The left-hand figure shows a histogram of the sample eigenvalues compared with the theoretical distribution and the ad-hoc fit. The right-hand figure shows the theoretical optimal shrinkage and the one obtained from the fit. The agreement is barely satisfactory, in particular the shrinkage from the fit is non-monotonic. The dots show the oracle estimator $\xi_k = \mathbf{v}_k^T \mathbf{C} \mathbf{v}_k$ computed within the same simulation.

given density. As a consequence the approximate estimator generated by such an ansatz is typically non-monotonic, whereas the exact shrinkage function should be.[2]

The Case of an Unbounded Support

On some real datasets, such as financial time series, it is hard to detect a clear boundary between bulk eigenvalues and the large outliers. In this case one may suspect that the distribution of eigenvalues of the true covariance matrix $\mathbf{C}$ is itself unbounded. In that case, one may try a parametric fit for which the density of $\mathbf{C}$ extends to infinity. For example, if we suspect that the true distribution has a sharp left edge but a power-law right tail, we may choose to model $\rho_{\mathbf{C}}(\lambda)$ as a shifted half Student's t-distribution, i.e.

$$\rho_{\mathbf{C}}(\lambda) = \Theta(\lambda - \lambda_-) \frac{2}{a\sqrt{\pi\mu}} \frac{\Gamma\left(\frac{1+\mu}{2}\right)}{\Gamma\left(\frac{\mu}{2}\right)} \left(1 + \frac{(\lambda - \lambda_-)^2}{a^2 \mu}\right)^{-\frac{1+\mu}{2}}, \tag{19.71}$$

where $\Theta(\lambda - \lambda_-)$ indicates that the density is non-zero only for $\lambda > \lambda_-$, chosen to be the center of the Student's t-distribution. These densities do not have an upper edge, instead they fall off as $\rho(\lambda) \sim \lambda^{-\mu-1}$ for large λ. For integer values of the tail

[2] Although we have not been able to find a simple proof of this property, we strongly believe that it holds in full generality.

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

exponent μ, the Stieltjes transform $\mathfrak{g}_{\mathbf{C}}(z)$ can be computed analytically. For example for $\mu = 3$ we find

$$\mathfrak{g}_{\mu=3}(z) = \frac{\sqrt{3}\pi u^3 + 6au^2 + 9a^2\sqrt{3}\pi u + 36a^3 \log\left(-u/\sqrt{3a^2}\right) + 18a^3}{\sqrt{3}\pi\left(3a^2 + u^2\right)^2}, \tag{19.72}$$

where $u = z - \lambda_-$. Note that this Stieltjes transform has an essential singularity at $z = \lambda_-$ and a branch cut on the real axis from λ_- to $+\infty$ indicating that the density has no upper bound. For $\mu = 3$ both the mean and the variance of the eigenvalue density are finite. We thus fix $\lambda_- = 1 - 2\sqrt{3a^2}/\pi$ such that $\tau(\mathbf{C}) = 1$ and adjust a to obtain the desired variance given by $\tau(\mathbf{C}^2) - 1 = 3a^2(1 - (2/\pi)^2)$.

In cases like this one, where we have an analytic form for $\mathfrak{g}_{\mathbf{C}}(z)$ but no simple formula for its S-transform, we can numerically solve the subordination relation

$$\mathfrak{t}_{\mathbf{E}}(\zeta) = \mathfrak{t}_{\mathbf{C}}\left(\frac{\zeta}{1 + q\mathfrak{t}_{\mathbf{E}}(\zeta)}\right), \tag{19.73}$$

with $\mathfrak{t}_{\mathbf{C}}(\zeta) = \zeta\mathfrak{g}_{\mathbf{C}}(\zeta) - 1$, using an efficient numerical fixed point equation solver. Most of the time a simple iteration would find the fixed point, but for some values of ζ and q it is sometimes difficult to find an initial condition for the iteration to converge so it is better to use a robust fixed point solver.

Let us end on a technical remark: for unbounded densities, $\mathfrak{g}(z)$ is not analytic at $z = \infty$, which does not conform to some hypotheses made throughout the book. Intuitively, there is no longer any clear distinction between bulk eigenvalues and outliers. For a fixed value of N, and for sufficiently large λ, the distance between two successive eigenvalues will at some point become much larger than $1/N$. Fortunately, the very same RIE formula holds both for bulk and for outlier eigenvalues, so we can close our eyes and safely apply Eq. (19.27) for unbounded densities as well.

19.5.2 Kernel Methods

Another approach to compute the Stieltjes and/or the T-transform on the real axis is to work directly with the discrete eigenvalues λ_k of $\mathbf{E}$. As stated earlier we cannot simply evaluate the discrete $g_N(z)$ at a point $z = \lambda_k$ because $g_N(z)$ is infinite precisely at the points $z \in \{\lambda_k\}$; this is the reason why $g_N(z)$ does not converge to the limiting $\mathfrak{g}_{\mathbf{E}}(z)$ on the support of $\rho(\lambda)$.

The idea here is to generalize the standard kernel method to estimate continuous densities from discrete data. Having observed a set of N eigenvalues $[\lambda_k]_{k\in(1,N)}$, a smooth estimator of the density is constructed as

$$\rho_{\mathrm{s}}(x) := \frac{1}{N}\sum_{k=1}^{N} K_{\eta_k}(x - \lambda_k), \tag{19.74}$$

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

where K_η is some adequately chosen kernel of width η (possibly k-dependent), normalized such that

$$\int_{-\infty}^{+\infty} \mathrm{d}u\, K_\eta(u) = 1, \tag{19.75}$$

such that

$$\int_{-\infty}^{+\infty} \mathrm{d}x\, \rho_{\mathrm{s}}(x) = 1. \tag{19.76}$$

A standard choice for K is a Gaussian distribution, but we will discuss more appropriate choices for the Stieltjes transform below.

Now, let us similarly define a smoothed Stieltjes transform as

$$\mathfrak{g}_{\mathrm{s}}(z) := \frac{1}{N}\sum_{k=1}^{N} \mathfrak{g}_{K,\eta_k}(z - \lambda_k), \tag{19.77}$$

where $\mathfrak{g}_{K,\eta}$ is the Stieltjes transform of the *kernel* K_η, treated as a density:

$$\mathfrak{g}_{K,\eta}(z) := \int_{-\infty}^{+\infty} \mathrm{d}u\, \frac{K_\eta(u)}{z-u}; \qquad \mathrm{Im}(z) \neq 0. \tag{19.78}$$

Note that since $\mathrm{Im}\, \mathfrak{g}_{K,\eta}(x - \mathrm{i}0^+) = \mathrm{i}\pi K_\eta(x)$, one immediately concludes that

$$\mathrm{Im}\, \mathfrak{g}_{\mathrm{s}}(x - \mathrm{i}0^+) = \mathrm{i}\pi\rho_{\mathrm{s}}(x) \tag{19.79}$$

for any smoothing kernel K_η. Hence $\mathfrak{g}_{\mathrm{s}}(z)$ is the natural generalization of smoothed densities for Stieltjes transforms. Correspondingly, the real part of the smoothed Stieltjes is the Hilbert transform (up to a π factor) of the smoothed density, i.e.

$$\mathfrak{h}_{\mathrm{s}}(x) := \mathrm{Re}\, \mathfrak{g}_{\mathrm{s}}(x - \mathrm{i}0^+) = \fint_{-\infty}^{\infty} \mathrm{d}\lambda \frac{\rho_{\mathrm{s}}(\lambda)}{x-\lambda}. \tag{19.80}$$

Two choices for the kernel K_η are specially interesting. One is the Cauchy kernel:

$$K_\eta^{\mathrm{C}}(u) := \frac{1}{\pi}\frac{\eta}{u^2+\eta^2}, \tag{19.81}$$

from which one gets

$$\mathfrak{g}_{K^{\mathrm{C}},\eta}(z) = \frac{1}{z \pm \mathrm{i}\eta}, \qquad \pm = \mathrm{sign}\,(\mathrm{Im}(z))\,. \tag{19.82}$$

Hence, in this case, we find that the smoothed Stieltjes transform we are looking for is nothing but the discrete Stieltjes transform computed with a k-dependent width η_k:

$$\mathfrak{g}_{\mathrm{s}}^{\mathrm{C}}(z) := \frac{1}{N}\sum_{k=1}^{N} \frac{1}{z - \lambda_k - \mathrm{i}\eta_k}, \qquad \mathrm{Im}(z) < 0, \tag{19.83}$$

which we can now safely compute numerically on the real axis, i.e. when $z = x - \mathrm{i}0^+$, and plug in the corresponding formulas for the RIE estimator $\xi(\lambda)$.

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

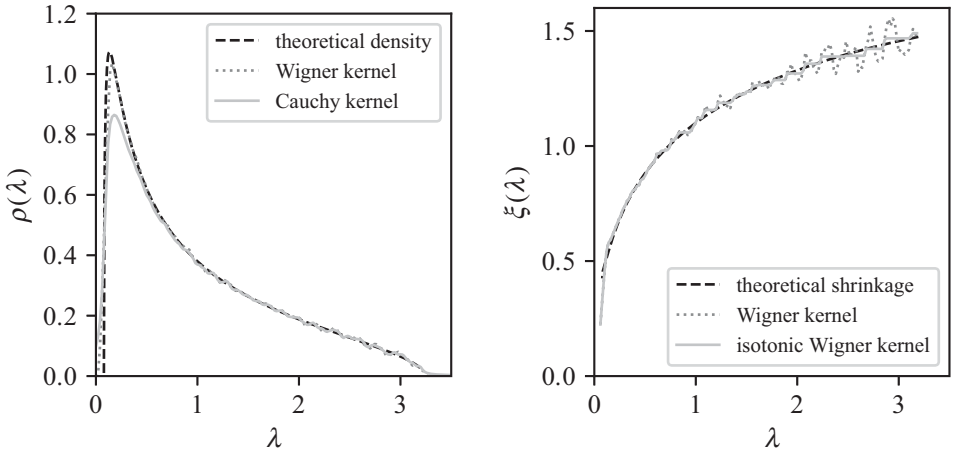

Figure 19.4 Non-parametric kernel methods applied to the same problem as in Figure 19.3. An approximation of $\mathfrak{g}_{\mathbf{E}}(\lambda - \mathrm{i}0^+)$ is computed with the Cauchy kernel (19.82) and the Wigner kernel (19.85) both with $\eta_k = \eta = N^{-1/2}$. (left) We compare the two smoothed densities with the theoretical one. Both are quite good but the Wigner kernel is better where the density changes rapidly. (right) From the smoothed Stieltjes transforms we compute the shrinkage function for both methods. Only the result of the Wigner kernel is shown (the Cauchy kernel is comparable albeit slightly worse). Kernel methods give non-monotonic shrinkage functions which can be easily rectified using an isotonic regression (19.86), which improves the agreement with the theoretical curve.

Another interesting choice for numerical applications is the semi-circle "Wigner kernel", which has sharp edges. To wit,

$$K_\eta^{\mathrm{W}}(u) = \frac{\sqrt{4\eta^2 - u^2}}{2\pi\eta^2} \quad \text{for} \quad -2\eta \le u \le 2\eta, \tag{19.84}$$

and 0 when $|u| > 2\eta$. In this case, we obtain

$$\mathfrak{g}_{\mathrm{s}}^{\mathrm{W}}(z) := \frac{1}{N}\sum_{k=1}^{N} \frac{z - \lambda_k}{2\eta_k^2}\left(1 - \sqrt{1 - \frac{4\eta_k^2}{(z-\lambda_k)^2}}\right). \tag{19.85}$$

Figure 19.4 gives an illustration of the kernel method using both the Cauchy kernel and the Wigner kernel, with $\eta_k = \eta = N^{-1/2}$.

We end this section with two practical implementation points regarding the kernel methods.

1 Since the optimal RIE estimator $\xi(\lambda)$ should be monotonic in λ, one should rectify possibly non-monotonic numerical estimators using an isotonic regression. The isotonic regression $\hat{y}_k$ of some data y_k is given by

$$\hat{y}_k = \operatorname{argmin} \sum_{k=1}^{T} \left(\hat{y}_k - y_k\right)^2 \quad \text{with} \quad \hat{y}_1 \le \hat{y}_2 \le \cdots \le \hat{y}_{T-1} \le \hat{y}_T. \tag{19.86}$$

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

It is the monotonic sequence that is the closest (in the least-square sense) to the original data.

2 In most situations, we are interested in reconstructing the optimal RIE matrix $\Xi = \sum \xi(\lambda_k)\mathbf{v}_k\mathbf{v}_k^T$ and hence we need to evaluate the shrinkage function $\xi(\lambda)$ precisely at the sample eigenvalues $\{\lambda_k\}$. We have found empirically that excluding the point λ_k itself from the kernel estimator consistently gives better results than including it. For example, in the Cauchy case, one should compute

$$g_s^C(\lambda_\ell - i0^+) \approx \frac{1}{N-1} \sum_{\substack{k=1 \\ k\neq\ell}}^{N} \frac{1}{\lambda_\ell - \lambda_k - i\eta_k}, \tag{19.87}$$

when estimating Eq. (19.48).

19.6 Validation and RIE

The idea of validation to determine the RIE is to compute the eigenvectors $\mathbf{v}_i$ of $\mathbf{E}$ the SCM of a *training set* and compute their unbiased variance of a different dataset: the *validation set*. More formally, this is written

$$\xi_\times(\lambda_i) := {\mathbf{v}_i}^T \mathbf{E}' \mathbf{v}_i, \tag{19.88}$$

where $\mathbf{E}'$ is the validation SCM. The training set is also called the *in-sample* data and the validation set the *out-of-sample* data.

In practical applications, we have typically a single dataset that needs to be split into a training and a validation set. If we are not too worried about temporal order, any block of the data can serve as the validation set. In *K-fold cross-validation*, the data is split into K blocks, one block is the validation set and the union of the $K-1$ others serves as the training set. The procedure is then repeated successively choosing the K possible validation sets, see Figure 19.5.

In the following, we will assume that the true covariance matrix $\mathbf{C}$ is the same on both datasets so that $\mathbf{E} = \mathbf{C}^{\frac{1}{2}}\mathbf{W}\mathbf{C}^{\frac{1}{2}}$ and $\mathbf{E}' = \mathbf{C}^{\frac{1}{2}}\mathbf{W}'\mathbf{C}^{\frac{1}{2}}$, where $\mathbf{W}'$ is independent from $\mathbf{W}$.

Expanding over the eigenvectors $\mathbf{v}'_j$ of $\mathbf{E}'$, we get

$$\xi_\times(\lambda_i) = \sum_{k=1}^{N} \left({\mathbf{v}_i}^T \mathbf{v}'_k\right)^2 \lambda'_k \tag{19.89}$$

or, in the large N limit and using the definition of Ψ given in Eq. (19.6),

$$\xi_\times(\lambda) \underset{N\to\infty}{\longrightarrow} \int \rho_{\mathbf{E}'}(\lambda')\Psi(\lambda,\lambda')\lambda' \, d\lambda'. \tag{19.90}$$

Now, there is an *exact* relation between Ψ and Φ, which reads

$$\Psi(\lambda,\lambda') = \frac{1}{N} \sum_{j=1}^{N} \Phi(\lambda,\mu_j)\Phi(\lambda',\mu_j) \tag{19.91}$$

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

or, in the continuum limit,

$$\Psi(\lambda,\lambda') = \int \rho_{\mathbf{C}}(\mu)\Phi(\lambda,\mu)\Phi(\lambda',\mu)\,\mathrm{d}\mu. \tag{19.92}$$

Intuitively, this relation can be understood as follows: we expect that, from the very definition of Φ, the eigenvectors of $\mathbf{E}$ and $\mathbf{E}'$ can be written as

$$\mathbf{v}_i = \frac{1}{\sqrt{N}}\sum_{j=1}^{N} \varepsilon_{ij}\sqrt{\Phi(\lambda_i,\mu_j)}\,\mathbf{u}_j; \qquad \mathbf{v}'_k = \frac{1}{\sqrt{N}}\sum_{\ell=1}^{N} \varepsilon'_{k\ell}\sqrt{\Phi(\lambda'_k,\mu_\ell)}\,\mathbf{u}_\ell, \tag{19.93}$$

where $\mathbf{u}_j$ are the eigenvectors of $\mathbf{C}$ and ε_{ij} are independent random variables of mean zero and variance one, such that

$$\mathbb{E}[\varepsilon_{ij}\varepsilon_{k\ell}] = \delta_{ik}\delta_{j\ell}, \qquad \mathbb{E}[\varepsilon'_{ij}\varepsilon'_{k\ell}] = \delta_{ik}\delta_{j\ell}, \qquad \mathbb{E}[\varepsilon_{ij}\varepsilon'_{k\ell}] = 0. \tag{19.94}$$

This so-called ergodic assumption can be justified from considerations about the Dyson Brownian motion of eigenvectors, see Eq. (9.10), but this goes beyond the scope of this book. In any case, if we now compute $\mathbb{E}[(\mathbf{v}_i^T\mathbf{v}'_k)^2]$ using the ergodic assumption and remembering that $\mathbf{u}_j^T\mathbf{u}_\ell = \delta_{j\ell}$, we find

$$N\mathbb{E}[(\mathbf{v}_i^T\mathbf{v}'_k)^2] = \frac{1}{N}\sum_{j=1}^{N}\Phi(\lambda_i,\mu_j)\Phi(\lambda'_k,\mu_j), \tag{19.95}$$

which is precisely Eq. (19.91).

Injecting Eq. (19.91) into Eq. (19.89), we thus find

$$\begin{aligned}\xi_\times(\lambda) &= \frac{1}{N^2}\sum_{k=1}^{N}\left(\sum_{j=1}^{N}\Phi(\lambda,\mu_j)\Phi(\lambda'_k,\mu_j)\right)\lambda'_k \\ &= \frac{1}{N^2}\sum_{j=1}^{N}\Phi(\lambda,\mu_j)\left(\sum_{k=1}^{N}\Phi(\lambda'_k,\mu_j)\lambda'_k\right).\end{aligned} \tag{19.96}$$

The last term in parenthesis can be computed by using the very definition of $\mathbf{E}'$:

$$\begin{aligned}\sum_{k=1}^{N}\Phi(\lambda'_k,\mu_j)\lambda'_k &\equiv N{\mathbf{u}_j}^T\mathbf{E}'\mathbf{u}_j \\ &= N\mathbf{C}^{\frac{1}{2}}\mathbf{u}_j^T\mathbf{W}'\mathbf{C}^{\frac{1}{2}}\mathbf{u}_j = N\mu_j\mathbf{u}_j^T\mathbf{W}'\mathbf{u}_j.\end{aligned} \tag{19.97}$$

Now, the idea is that since $\mathbf{W}'$ is independent of $\mathbf{C}$, averaging over any small interval of μ_j will amount to replacing ${\mathbf{u}_j}^T\mathbf{W}'\mathbf{u}_j$ by its average over randomly oriented vectors $\mathbf{u}$, which is equal to unity:

$$\mathbb{E}[\mathbf{u}^T\mathbf{W}'\mathbf{u}] = \tau(\mathbf{W}'\mathbf{u}\mathbf{u}^T) = \tau(\mathbf{W}')\tau(\mathbf{u}\mathbf{u}^T) = 1. \tag{19.98}$$

Hence, from Eq. (19.96) we finally obtain

$$\xi_\times(\lambda) = \frac{1}{N}\sum_{j=1}^{N}\Phi(\lambda,\mu_j)\mu_j \underset{N\to\infty}{\longrightarrow} \int \rho_{\mathbf{C}}(\mu)\,\Phi(\lambda,\mu)\,\mu\,\mathrm{d}\mu, \tag{19.99}$$

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

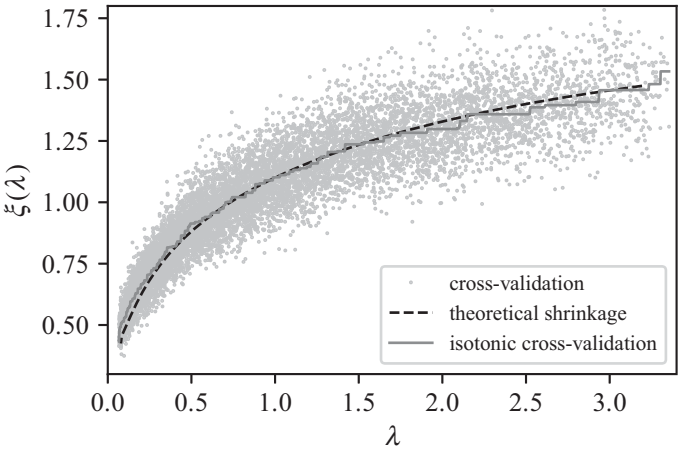

Figure 19.5 Shrinkage function $\xi(\lambda)$ computed for the same problem as in Figure 19.3, now using cross-validation. The dataset is divided into $K = 10$ blocks of equal length. For each block, we compute the $N = 1000$ eigenvalues λ_i^b and eigenvectors $\mathbf{v}_i^b$ of the sample covariance matrix using the rest of the data (of new length $9T/10$), and compute $\xi_\times(\lambda_i^b) := \mathbf{v}_i^{b\,T}\mathbf{E}'\mathbf{v}_i^b$, with $\mathbf{E}'$ the sample covariance matrix of the considered block. The dots correspond to the 10×1000 pairs $(\lambda_i^b, \xi_\times(\lambda_i^b))$. The full line is an isotonic regression through the dots. The procedure has a slight bias as we in fact compute the optimal shrinkage for a value of q equal to $q_\times = 10N/9T$, but otherwise the agreement with the optimal curve is quite good.

which precisely coincides with the definition of the optimal non-linear shrinkage function $\xi(\lambda)$, see Eq. (19.27).

This result is very interesting and indicates that one can approximate $\xi(\lambda)$ by considering the quadratic form between the eigenvectors of a given realization of $\mathbf{C}$ – say $\mathbf{E}$ – and another realization of $\mathbf{C}$ – say $\mathbf{E}'$ – even if the two empirical matrices are characterized by different values of the ratio N/T. This method is illustrated in Figure 19.5.

Bibliographical Notes

- For a recent review on the subject of cleaning noisy covariance matrices and RIE, see
  - J. Bun, J.-P. Bouchaud, and M. Potters. Cleaning correlation matrices. *Risk magazine*, 2016,
  - J. Bun, J.-P. Bouchaud, and M. Potters. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.
- The original work of Ledoit and Péché and its operational implementation, see
  - O. Ledoit and S. Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011,
  - O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012,

https://doi.org/10.1017/9781108768900.023 Published online by Cambridge University Press

– O. Ledoit and M. Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 2017.

- Rotationally invariant estimators for general additive and multiplicative models:
  – J. Bun, R. Allez, J.-P. Bouchaud, and M. Potters. Rotational invariant estimator for general noisy matrices. *IEEE Transactions on Information Theory*, 62:7475–7490, 2016.
- Rotationally invariant estimators for outliers:
  – J. Bun and A. Knowles. An optimal rotational invariant estimator for general covariance matrices. Unpublished, 2016. preprint available on researchgate.net,,

  see also Bun et al. [2017].
- Overlaps between the eigenvectors of correlated matrices:
  – J. Bun, J.-P. Bouchaud, and M. Potters. Overlaps between eigenvectors of correlated random matrices. *Physical Review E*, 98:052145, 2018.
- Rotationally invariant estimators for cross-correlation matrices:
  – F. Benaych-Georges, J.-P. Bouchaud, and M. Potters. Optimal cleaning for singular values of cross-covariance matrices. *preprint arXiv:1901.05543*, 2019.