Math 182: Hidden Data in Random Matrices

Todd Kemp

Contents

CHAPTER 1

Principal Component Analysis

Consider a high dimensional data set

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^m$$

where by *high-dimensional*, typically we mean $m \gg N$: the ambient dimension is much larger than the number of data points. (A typical example would be $N = 1000$ points in $m = 10^6$ dimensions.) There may be no natural division of variables into *predictor* and *response* variables. Instead, our hope is that there is a (probably *much*) lower-dimensional affine subspace of $\mathbb{R}^m$ where, in some sense, all the data "approximately lives". More precisely: our model is that the "true" data points $\mathbf{t}_j$ all live in some lower-dimensional affine subspace, and the measured data $\mathbf{x}_j$ are samples of $\mathbf{X}_j = \mathbf{t}_j + \mathbf{Z}_j$, where the $\mathbf{Z}_j$ are random vectors modeling noise and error. We could attempt to setup a MLE (Maximum Likelihood Estimator) based on this model, but it is a little unclear what the unknown parameters are; how do we describe the desired affine subspace?

1.1. Dimension Reduction via Affine Projection

To begin, let's make the term *affine subspace* precise.

DEFINITION 1.1. *Fix positive integers $1 \leq d \leq m$. An* **affine subspace** *$\mathcal{A} \subseteq \mathbb{R}^m$ is a set of vectors of the form*

$$\mathcal{A} = \boldsymbol{\mu} + V$$

for some fixed vector $\boldsymbol{\mu} \in \mathbb{R}^m$ and some subspace $V \subseteq \mathbb{R}^m$ with $\dim(V) = d$. That is: $\mathcal{A} = \{\boldsymbol{\mu} + \mathbf{v} \colon \mathbf{v} \in V\}$.

EXAMPLE 1.2. In $\mathbb{R}^2$, the graphs of all lines $y = mx + b$, together with vertical lines $x = a$, for $a, b, m \in \mathbb{R}$, are the affine subspaces.

EXAMPLE 1.3. If $\mathcal{A}$ is an affine subspace in the form $\mathcal{A} = \boldsymbol{\mu} + V$, and $\mathbf{v_0} \in V$, then we can also write $\mathcal{A} = (\boldsymbol{\mu} + \mathbf{v}_0) + V$. Indeed:

- If $\mathbf{w} \in V$, then $(\boldsymbol{\mu} + \mathbf{v}_0) + \mathbf{w} = \boldsymbol{\mu} + (\mathbf{v}_0 + \mathbf{w}) \in \boldsymbol{\mu} + V$; this shows $(\boldsymbol{\mu} + \mathbf{v}_0) + V \subseteq \mathcal{A}$.
- Conversely, if $\mathbf{a} \in \mathcal{A} = \boldsymbol{\mu} + V$, then there is some $\mathbf{v} \in V$ such that $\mathbf{a} = \boldsymbol{\mu} + \mathbf{v} = (\boldsymbol{\mu} + \mathbf{v}_0) + (\mathbf{v} - \mathbf{v}_0)$. Since $\mathbf{v} - \mathbf{v}_0 \in V$, this shows $\mathbf{a} \in (\boldsymbol{\mu} + \mathbf{v}_0) + V$, and so $\mathcal{A} \subseteq (\boldsymbol{\mu} + \mathbf{v}_0) + V$.

All this is to show that we cannot talk about *the* vector $\boldsymbol{\mu}$ that translates V away from $\mathbf{0}$ to define $\mathcal{A}$; rather, there is a whole family of such translation vectors (indexed by V). Put another way: we could define an *affine subspace* $\mathcal{A}$ to be a subset with the property that $\mathcal{A} - \mathcal{A}$ is a subspace; i.e. there is a subspace V such that the difference of any two elements in $\mathcal{A}$ is in V.

As Example 1.3 shows, the vector used to translate the subspace away from $\mathbf{0}$ isn't unique; on the other hand, we do need one such vector to specify the affine subspace. What other information do we need? How do we fully describe (non-uniquely) an affine subspace? Let's answer that with a proposition.

PROPOSITION 1.4. *Any affine subspace $\mathcal{A}$ of dimension d is described in the form*

$$\mathcal{A} = \boldsymbol{\mu} + \mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$$

for some fixed vector $\boldsymbol{\mu}$ and some linearly independent set $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ of d vectors. Moreover, if there is another such representation

$$\mathcal{A} = \boldsymbol{\nu} + \mathrm{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_d\}$$

then $\{\mathbf{u}_1, \ldots, \mathbf{u}_d\}$ is a basis of the same subspace $V = \mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$, and $\boldsymbol{\mu} - \boldsymbol{\nu} \in V$.

(The proof of this proposition is an exercise in basic linear algebra, following the discussion in Example 1.3.) What we learn from this is: to specify an affine subspace of dimension d, we need $d + 1$ vectors: a translation vector $\boldsymbol{\mu}$, and a linearly independent set $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ spanning the subspace part. We will utilize the large amount of freedom in choosing these vectors to be a bit more restrictive, and always choose an **orthonormal basis** for the subspace. Let's now remind ourselves of a few basic facts about orthonormal bases of subspaces.

PROPOSITION 1.5. *Let $V \subseteq \mathbb{R}^m$ be a subspace, and let $\{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d\}$ be an orthonormal basis for V. Let Q be the $m \times d$ matrix with $\hat{\mathbf{u}}_j$ as columns:*

$$Q = \begin{bmatrix} | & | & & | \\ \hat{\mathbf{u}}_1 & \hat{\mathbf{u}}_2 & \cdots & \hat{\mathbf{u}}_d \\ | & | & & | \end{bmatrix}.$$

Then $Q^\top Q = I_d$ is the $d \times d$ identity matrix, while $QQ^\top$ the $m \times m$ matrix of the orthogonal projection P_V onto the subspace V.

PROOF. By definition, $[Q^\top Q]_{ij} = \hat{\mathbf{u}}_i^\top \hat{\mathbf{u}}_j = \hat{\mathbf{u}}_i \cdot \hat{\mathbf{u}}_j = \delta_{ij}$ since they form and orthonormal set; this shows $Q^\top Q = I_d$. In the other direction: for any vector $\mathbf{w} \in \mathbb{R}^m$, $[Q^\top \mathbf{w}]_j = \hat{\mathbf{u}}_j^\top \mathbf{w} = \hat{\mathbf{u}}_j \cdot \mathbf{w}$; thus, by the definition of matrix multiplication,

$$QQ^\top \mathbf{w} = Q(Q^\top \mathbf{w}) = \sum_{j=1}^{d} (\hat{\mathbf{u}}_j \cdot \mathbf{w}) \hat{\mathbf{u}}_j$$

which (because of the orthonormality) is the orthogonal projection of $\mathbf{w}$ onto $\mathrm{span}\{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d\} = V$. □

Hence, we can alternatively think if specifying a d-dimensional affine subspace of $\mathbb{R}^m$ with two objections: (1) a vector $\boldsymbol{\mu} \in \mathbb{R}^m$, and (2) an $m \times d$ matrix Q satisfying $Q^\top Q = I_d$. Given such a pair $(\boldsymbol{\mu}, Q)$, the affine subspace is then the $\boldsymbol{\mu}$-shift of the column space $\mathrm{Col}(Q)$; alternatively, it is the $\boldsymbol{\mu}$-shift of the range of the orthogonal projection $QQ^\top$.

REMARK 1.6. In the same spirit, to be minimal, we could also assume that the vector $\boldsymbol{\mu}$ is orthogonal to V. After all, once one translation vector $\boldsymbol{\mu}_0$ has ben chosen, we note that the orthogonal projection $\mathbf{v} = QQ^\top \boldsymbol{\mu}_0$ is defined by the property that $\boldsymbol{\mu}_0 - \mathbf{v} \perp V$; but from Proposition 1.4, the vector $\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \mathbf{v}$ may just as well be used as a translation vector for the affine subspace. (We do not, however, have the freedom to normalize the vector $\boldsymbol{\mu}$; in general, the one orthogonal to V just constructed will be the *shortest* possible choice.)

However, as we will shortly see, it will be more convenient to give the freedom to add vectors (in V) to the translation vector to make a canonical choice in our data analysis framework. So we will generally not place orthogonality constraints on $\boldsymbol{\mu}$ with respect to V.

1.2. Least Squares, and the Best Translation Vector

Returning to our setup, with more precise language: we have data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in $\mathbb{R}^m$. Our goal is to find *the best approximating d-dimensional affine subspace* for this data. As with linear regression, we *could* approach this by setting up an appropriate Gaussian noise model and developing the corresponding MLE; as with the calculation in the regression case, it is straightforward to see where that will lead: to a *least squares approximation*. We will therefore set it up that way from the start.

In seeking the "best fit affine subspace", we will look for a pair $(\boldsymbol{\mu}, Q)$ to specify it: a translation vector and an orthonormal basis (written as an $m \times d$ matrix Q satisfying $Q^\top Q = I_d$). *These will not be unique!* Any such pair can be replaced with infinitely many others by translating $\boldsymbol{\mu}$ by a vector in $\mathrm{Col}(Q)$, and by replacing the given orthonormal basis columns in Q with another one (which boils down to multiplying Q on the right by a $d \times d$ orthogonal matrix). All this means is that the minimization problem will not have a unique solution; we will choose a minimizer that is "canonical" (essentially that is easiest, and most meaningful, to compute).

To get started: the points is to find an affine subspace, specified by some $(\boldsymbol{\mu}, Q)$, such that

$$\text{for } 1 \leq j \leq N, \quad \mathbf{x}_j - \boldsymbol{\mu} \text{ is close to the subspace } \mathrm{Col}(Q).$$

That is: we should be able to find, for each of the data points $\mathbf{x}_j$, a vector $\boldsymbol{\beta}_j \in \mathbb{R}^d$, so that

$$\mathbf{x}_j \approx \boldsymbol{\mu} + Q\boldsymbol{\beta}_j.$$

This leads us to a precise definition of what *best fit* should mean.

DEFINITION 1.7. *Given data points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ in $\mathbb{R}^m$, a d-dimensional affine subspace specified by a pair $(\boldsymbol{\mu}_0, Q_0)$ is said to be a* **best fit** *for the data if the function*

$$\Phi^d(\boldsymbol{\mu}, Q, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) = \sum_{j=1}^{N} \|\mathbf{x}_j - (\boldsymbol{\mu} + Q\boldsymbol{\beta}_j)\|^2$$

(defined for $\boldsymbol{\mu} \in \mathbb{R}^m$, $Q \in \mathbb{M}_{m\times d}$ satisfying $Q^\top Q = I_d$, and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N \in \mathbb{R}^d$) achieves its minimum at $(\boldsymbol{\mu}, Q) = (\boldsymbol{\mu}_0, Q_0)$ (for some choices of $\boldsymbol{\beta}_j$).

This now becomes something like a linear regression problem: we are looking for the best affine function $\boldsymbol{\beta} \mapsto \boldsymbol{\mu}_0 + Q_0\boldsymbol{\beta}$ to fit the data. The catch is: unlike in linear regression, the "predictor" variables $\boldsymbol{\beta}_j$ are not given; they must be discovered by the least squares minimization. This makes the problem computationally harder, as we will see.

REMARK 1.8. It is crucial to note that the dimension d of the best fit affine subspace *must be specified from the start*. We cannot include d as a parameter in the minimization. Indeed, if we did, the minimum would always be achieved with $d = m$, where we can take $\boldsymbol{\mu}_0 = \mathbf{0}$, $Q_0 = I_m$, and $\boldsymbol{\beta}_j = \mathbf{x}_j$, achieving the least-squares minimum of 0. (That is: the best m-dimensional approximation of the data is the original data, since it was m-dimensional to start with!) This, of course, defeats the purpose: the name of the game here is dimension *reduction*.

The question of how to choose the "right" d is the really subtle part, which we will spend much more time trying to understand later.

We will fully solve this minimization problem in this chapter, by leveraging all of the linear algebra we've discussed. As a warm-up, we can begin by partially solving the minimization, to find a best fit $\boldsymbol{\mu}_0$ irrespective of the other parameters.

THEOREM 1.9. *A global minimum of Φ^d occurs at at a point with translation vector*

$$\boldsymbol{\mu}_0 = \overline{\mathbf{x}}_N = \frac{1}{N}\sum_{j=1}^{N} \mathbf{x}_j.$$

PROOF. The function $\Phi^d(\boldsymbol{\mu}, Q, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ is smooth (in fact it is a quadratic polynomial) in all the parameters $\boldsymbol{\mu}, Q, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$. Hence, we should seek critical points to find local extrema. (A complete analysis would then also show that the local minimum is global by considering the behavior of the function as the parameters grow without bound; in the interest of staying on task, we will omit this less interesting argument.)

We therefore take the directional derivative in the $\boldsymbol{\mu}$ direction (holding Q and $\boldsymbol{\beta_j}$ fixed); that is, we seek a $\boldsymbol{\mu}_0$ satisfying

$$0 = \frac{d}{dt}\Phi^d(\boldsymbol{\mu}_0 + t\mathbf{z}, Q, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)\Big|_{t=0}$$

for each $\mathbf{z} \in \mathbb{R}^m$. Note that

$$\|\mathbf{x}_j - (\boldsymbol{\mu}_0 + t\mathbf{z} + Q\boldsymbol{\beta}_j)\|^2 = \|\mathbf{x}_j - \boldsymbol{\mu}_0 - Q\boldsymbol{\beta}_j\|^2 + 2t\,(\mathbf{x}_j - \boldsymbol{\mu}_0 - Q\boldsymbol{\beta}_j)\cdot\mathbf{z} + t^2\|\mathbf{z}\|^2$$

and the derivative of this (with respect to t) at $t = 0$ is just $2(\mathbf{x}_j - \boldsymbol{\mu}_0 - Q\boldsymbol{\beta}_j)\cdot\mathbf{z}$. Hence, adding up terms, the critical point equation becomes

$$0 = \sum_{j=1}^{N} 2(\mathbf{x}_j - \boldsymbol{\mu}_0 - Q\boldsymbol{\beta}_j)\cdot\mathbf{z}, \qquad \text{for all } \mathbf{z} \in \mathbb{R}^m.$$

This shows that the vector

$$\sum_{j=1}^{N}(\mathbf{x}_j - \boldsymbol{\mu}_0 - Q\boldsymbol{\beta}_j) = N\overline{\mathbf{x}}_N - N\boldsymbol{\mu}_0 - Q(\sum_{j=1}^{N}\boldsymbol{\beta}_j)$$

is orthogonal to *all* vectors $\mathbf{z} \in \mathbb{R}^m$, which means it must be the zero vector $\mathbf{0}$. Hence, a critical value for the minimizing $\boldsymbol{\mu}_0$ is achieved at

$$\boldsymbol{\mu}_0 = \overline{\mathbf{x}}_N - \frac{1}{N}Q(\sum_{j=1}^{N}\boldsymbol{\beta}_j) = \overline{\mathbf{x}}_N - Q(\overline{\boldsymbol{\beta}}_N)$$

where $\overline{\boldsymbol{\beta}}_N = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\beta}_j$ (here we have used the linearity of matrix multiplication to put the $\frac{1}{N}$ inside).

It appears from this that the minimizing $\boldsymbol{\mu}_0$ depends on the minimizing $\boldsymbol{\beta}_j$'s; however, as discussed above, the minimizers are highly non-unique. In particular, since $Q(\overline{\boldsymbol{\beta}}_N)$ is (by definition) in the Column space of Q, the affine subspace specified by $(\overline{\mathbf{x}}_N, Q)$ is *the same* as the one specified by $(\overline{\mathbf{x}}_N - Q(\overline{\boldsymbol{\beta}}_N), Q)$ (cf. Proposition 1.4). Hence, we have shown that, independent of the minimizers in the Q and $\boldsymbol{\beta}_j$ variables (and the dimension d), the minimum of Φ^d is achieved with translation vector $\boldsymbol{\mu}_0 = \overline{\mathbf{x}}_N$, as stated.

□

1.3. The Best Fit Subspace vs. Linear Regression

Before completing the general minimization problem of Definition 1.7, let's take a small detour to compare this approach to linear regression. Recall that linear regress deals with data of the form $(\mathbf{x}_j, y_j)_{j=1}^N$, where the expectation is that $\mathbf{x}_j$ is a predictor of y_j: i.e. there is some affine relationship $y_j \approx \mathbf{w}_0 \cdot \mathbf{x}_j + b_0$. Th model of random noise used in that setting let do the minimization problem

$$(\mathbf{w}_0, b_0) = \operatorname*{argmin}_{\mathbf{w},b} \sum_{j=1}^N |y_j - (\mathbf{w} \cdot \mathbf{x}_j + b)|^2 \tag{1.1}$$

While somewhat similar in form to Definition 1.7, there is substantial differences, which fundamentally result from the coordinate y being treaded differently from the coordinates $\mathbf{x}$ in the vector $(\mathbf{x}, y)$. The regression minimization is to minimize the sum of squares of the *vertical displacements* (in the y-coordinate) of the points $(\mathbf{x}_j, y_j)$ from a codimension-1 affine subspace. Definition 1.7, on the other hand, treats all coordinates equally.

To understand a bit better what this means geometrically, let's consider the dimension 2 case, with best approximating $d = 1$ dimensional affine subspace. Here the data have the form $\mathbf{x}_j = (x_j, y_j)$. The minimization problem of Definition 1.7 is then to find minimizing $\boldsymbol{\mu} \in \mathbb{R}^2$; $Q \in \mathbb{M}_{2\times 1} = \mathbb{R}^2$ satisfying $Q^\top Q = I_1 = 1$, i.e. $Q = \hat{\mathbf{u}}$ is a unit vector; and "vectors" $\beta_j \in \mathbb{R}^1$. That is, we want to minimize

$$(\boldsymbol{\mu}, \hat{\mathbf{u}}, \beta_1, \ldots, \beta_N) \mapsto \sum_{j=1}^N \|\mathbf{x}_j - (\boldsymbol{\mu} + \hat{\mathbf{u}}\beta_j)\|^2.$$

We have already computed, in the last section, that an optimal translation vector is $\boldsymbol{\mu} = \overline{\mathbf{x}}_N$; i.e. we seek to minimize

$$(\hat{\mathbf{u}}, \beta_1, \ldots, \beta_N) \mapsto \sum_{j=1}^N \|(\mathbf{x}_j - \overline{\mathbf{x}}_N) - \hat{\mathbf{u}}\beta_j\|^2.$$

Let's denote the shifted data as $\mathbf{x}_j - \overline{\mathbf{x}}_N =: \mathbf{x}_j^\circ$. Now, suppose we have already found the optimal unit vector $\hat{\mathbf{u}}$; then we can find the optimal β_j easily, finding the critical values. Differentiating with respect to β_i eliminates all but one term in the sum, so we have

$$0 = \frac{\partial}{\partial \beta_i}\|\mathbf{x}_i^\circ - \hat{\mathbf{u}}\beta_i\|^2 = \frac{\partial}{\partial \beta_i}(\|\mathbf{x}_i\|^2 - 2\mathbf{x}_i^\circ \cdot \hat{\mathbf{u}}\,\beta_i + \|\hat{\mathbf{u}}\|^2\beta_i^2).$$

Since $\|\hat{\mathbf{u}}\| = 1$, this equation becomes

$$0 = -2\mathbf{x}_i^\circ \cdot \hat{\mathbf{u}} + 2\beta_i \quad \Rightarrow \quad \beta_i = \mathbf{x}_i^\circ \cdot \hat{\mathbf{u}}.$$

That is: the minimum least squares sum is

$$\sum_{j=1}^N \|\mathbf{x}_j^\circ - (\mathbf{x}_j^\circ \cdot \hat{\mathbf{u}})\hat{\mathbf{u}}\|^2.$$

The vector inside the sum should be familiar: $(\mathbf{x}_j^\circ \cdot \hat{\mathbf{u}})\hat{\mathbf{u}}$ is the *orthogonal projection* of $\mathbf{x}_j^\circ$ onto $\operatorname{span}(\hat{\mathbf{u}})$, and the different is thus

$$\mathbf{x}_j^\circ - (\mathbf{x}_j^\circ \cdot \hat{\mathbf{u}})\hat{\mathbf{u}} = \operatorname{Proj}_{\hat{\mathbf{u}}^\perp}(\mathbf{x}_j^\circ).$$

This now explains exactly what the optimal $\hat{\mathbf{u}}$ is, and how this minimization problem compares to linear regression. Let's summarize the conclusion as a proposition.

PROPOSITION 1.10. *With $m = 2$ dimensional data $(\mathbf{x}_j)_{j=1}^N$, the best fit $d = 1$ dimensional affine subspace (cf. Definition 1.7) is defined by translating the data by its sample mean $\mathbf{x}_j^\circ = \mathbf{x}_j - \overline{\mathbf{x}}_N$, and then finding the line (passing through the origin) that minimizes the (sum of squares) distances from the points $\mathbf{x}_j^\circ$ to the line* **orthogonally**. *(This is contrasted with linear regression, cf.* (1.1) *where the best fit line is defined by minimizing the sum of squares of the* **vertical** *displacements of the data to the line.)*

REMARK 1.11. In this special case, the process of finding the best fit 1-dimensional subspace is sometimes called *orthogonal regression* or *total regression*. While the general process of finding the best fit affine subspace of a given dimension doesn't quite conform to regression, orthogonal projections play a key role in general, as we will see going forward.

1.4. The Best Fit Coordinates $\boldsymbol{\beta}_j$

We can follow the calculations in the previous section (in the special case $m = 2$, $d = 1$) almost verbatim to compute the best-fit $\boldsymbol{\beta}_j$ in general. As above, set

$$\mathbf{x}_j^\circ := \mathbf{x}_j - \overline{\mathbf{x}}_N. \tag{1.2}$$

We have already shown that a best fit translation vector is $\boldsymbol{\mu}_0 = \overline{\mathbf{x}}_N$; thus, we wish to minimize

$$\Phi^d(\overline{\mathbf{x}}_N, Q, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N) = \sum_{j=1}^N \|\mathbf{x}_j^\circ - Q\boldsymbol{\beta}_j\|^2$$

over all possible coordinate vectors $\boldsymbol{\beta}_j$.

REMARK 1.12. We refer to the $\boldsymbol{\beta}_j$ as *coordinates*. If the vector $\mathbf{x}_j^\circ$ were actually in the subspace $\mathrm{Col}(Q)$, then it would have a unique expansion $\mathbf{x}_j^\circ = Q\boldsymbol{\beta}_j$ for some $\boldsymbol{\beta}_j \in \mathbb{R}^d$; this vector is then the **coordinate vector** of $\mathbf{x}_j^\circ$ in the basis $\{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d\}$ of columns of Q. In reality, $\mathbf{x}_j^\circ$ is not *in* the subspace, so it doesn't have coordinates in this basis, but we'll still refer to the $\boldsymbol{\beta}_j$ as coordinate vectors.

Fixing Q for now, the terms in the sum all decouple for different $\boldsymbol{\beta}_i$; hence, looking for a local maximum, we must solve the critical point equations

$$\frac{d}{dt}\|\mathbf{x}_i^\circ - Q(\boldsymbol{\beta}_i + t\boldsymbol{\gamma})\|^2\Big|_{t=0}, \quad \boldsymbol{\gamma} \in \mathbb{R}^d,\ 1 \le i \le N.$$

Expanding the summed norm,

$$\|\mathbf{x}_i^\circ - Q(\boldsymbol{\beta}_i + t\boldsymbol{\gamma})\|^2 = \|\mathbf{x}_i^\circ - Q\boldsymbol{\beta}_i\|^2 + t(\mathbf{x}_i^\circ - Q\boldsymbol{\beta}_i)\cdot Q\boldsymbol{\gamma} + t^2\|Q\boldsymbol{\gamma}\|^2.$$

Taking the derivative of this expression at $t = 0$, the equations become

$$0 = (\mathbf{x}_i^\circ - Q\boldsymbol{\beta}_i)\cdot Q\boldsymbol{\gamma} = Q^\top(\mathbf{x}_i^\circ - Q\boldsymbol{\beta}_i)\cdot\boldsymbol{\gamma}, \qquad \boldsymbol{\gamma} \in \mathbb{R}^d,\ 1 \le i \le N.$$

As we argued above: since the vector $Q^\top(\mathbf{x}_i^\circ - Q\boldsymbol{\beta}_i)$ is orthogonal to *all* vectors $\boldsymbol{\gamma} \in \mathbb{R}^d$, it must be the $\mathbf{0}$ vector. Hence, we have

$$\mathbf{0} = Q^\top(\mathbf{x}_i^\circ - Q\boldsymbol{\beta}_i) = Q^\top\mathbf{x}_i^\circ - Q^\top Q\boldsymbol{\beta}_i$$

and, using the fact that $Q^\top Q = I_d$, we finally conclude that

$$\boldsymbol{\beta}_i = Q^\top\mathbf{x}_i^\circ, \quad 1 \le i \le N.$$

Subbing this back into the function Φ^d, we see a pleasing conclusion.

COROLLARY 1.13. *For any fixed subspace $V \subseteq \mathbb{R}^m$ of dimension d, specified by an orthonormal basis matrix $Q \in \mathbb{M}_{m\times d}$, the minimum in problem of Definition 1.7 is achieved at*

$$\sum_{j=1}^{N} \|\mathbf{x}_j^\circ - QQ^\top \mathbf{x}_j^\circ\|^2 = \sum_{j=1}^{N} \|\mathrm{Proj}_{V^\perp}\mathbf{x}_j^\circ\|^2. \tag{1.3}$$

That is: for any data set, and *fixed* subspace, the least squares problem for a data set $\{\mathbf{x}_j\}$ relative to V is to *translate the data by its sample mean and then orthogonally project into V*. We might call this an **affine projection**: orthogonal projection of the data into an affine subspace (by first translating the affine subspace to $\mathbf{0}$ and the applying the projection). The quantity in (1.3) is a measure of how far the data are from this affine projection image.

REMARK 1.14. Now we can say more precisely what the $\boldsymbol{\beta}_j = Q^t op\mathbf{x}_j^\circ$ "coordinate vectors" really are: they are the coordinates of the *projected vector* $QQ^\top\mathbf{x}_j^\circ$ in the basis of columns of Q.

1.5. The Best Fit Subspace, and the Sample Covariance Matrix

Reiterating: we have reduced the least squares minimization problem in Definition 1.7 to the following. Given data $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in $\mathbb{R}^m$, with sample mean $\overline{\mathbf{x}}_N = \frac{1}{N}\sum_{j=1}^N \mathbf{x}_j$, centering the data $\mathbf{x}_j^\circ = \mathbf{x}_j - \overline{\mathbf{x}}_N$, we seek a global minimizer Q_0 of the following problem:

$$Q_0 = \underset{Q\in\mathbb{M}_{m\times d}, Q^\top Q = I_d}{\operatorname{argmin}} \sum_{j=1}^{N} \|\mathbf{x}_j^\circ - QQ^\top\mathbf{x}_j^\circ\|^2. \tag{1.4}$$

We arrived at this form by minimizing over the best translation vector $\boldsymbol{\mu}$, and coefficients $\boldsymbol{\beta}_j$ for the data in the (unknown) basis Q; these minimization problems were fairly straightforward using calculus, since the sets of possible $\boldsymbol{\mu}$ and $\boldsymbol{\beta}_j$ were *vector spaces*. We could therefore find the minimizers by considering linear perturbations (e.g. $\boldsymbol{\mu} + t\mathbf{z}$) and differentiating. The minimization problem for Q, however, is not so simple. The set of possible minimizers Q (those satisfying $Q^\top Q = I$) is not a linear space. (It is a nice geometric space, called a *symmetric space*, possessing a nice action of the Lie group $\mathrm{O}(d)$; this minimization problem could be approached therefore as a constrained optimization problem via calculus on manifolds, but that is now how we will approach it.)

To attack (1.4), we can make one quick simplification by expanding out the squared norm.

$$\|\mathbf{x}_j^\circ - QQ^\top\mathbf{x}_j^\circ\|^2 = \|\mathbf{x}_j^\circ\|^2 - 2\mathbf{x}_j^\circ \cdot QQ^\top\mathbf{x}_j^\circ + \|QQ^\top\mathbf{x}_j^\circ\|^2. \tag{1.5}$$

Notice that the last term is

$$\|QQ^\top\mathbf{x}_j^\circ\|^2 = QQ^\top\mathbf{x}_j^\circ \cdot QQ^\top\mathbf{x}_j^\circ = \mathbf{x}_j^\circ \cdot (QQ^\top)^\top QQ^\top\mathbf{x}_j^\circ.$$

Since $QQ^\top$ is an orthogonal projection, $(QQ^\top)^\top QQ^\top = QQ^\top$. (This is easy to compute directly: $(QQ^\top)^\top QQ^\top = ((Q^\top)^\top Q^\top)QQ^\top = Q(Q^\top Q)Q^\top = QIQ^\top = QQ^\top$.) Hence, the last two terms in (1.5) are the same (modulo a factor of -2), and so the expression is equal to

$$\|\mathbf{x}_j^\circ\|^2 - \mathbf{x}_j^\circ \cdot QQ^\top\mathbf{x}_j^\circ.$$

REMARK 1.15. The savvy reader will realize that the preceding calculation is exactly the same as the one showing that, for any L^2 random variable X, $\mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. Indeed, one can think of applying the projection $QQ^\top$ as a kind of (conditional) "expectation".

We can now rewrite the function being minimized as

$$Q \mapsto \sum_{j=1}^{N} \|\mathbf{x}_j^\circ\|^2 - \sum_{j=1}^{N} \mathbf{x}_j^\circ \cdot QQ^\top \mathbf{x}_j^\circ.$$

The first sum is an additive constant with respect to Q, and so it does not affect the minimizer. The second sum has a minus sign, and thus, (1.4) is equivalent to the following *maximization* problem:

$$Q_0 = \operatorname*{argmax}_{Q \in \mathbb{M}_{m\times d}, Q^\top Q = I_d} \sum_{j=1}^{N} \mathbf{x}_j^\circ \cdot QQ^\top \mathbf{x}_j^\circ. \tag{1.6}$$

It will be convenient to rewrite the terms in this sum. First, by definition,

$$\mathbf{x}_j^\circ \cdot QQ^\top \mathbf{x}_j^\circ = (\mathbf{x}_j^\circ)^\top QQ^\top \mathbf{x}_j^\circ = (Q^\top \mathbf{x}_j^\circ)^\top Q^\top \mathbf{x}_j^\circ = \|Q^\top \mathbf{x}_j^\circ\|^2$$

so

$$Q_0 = \operatorname*{argmax}_{Q \in \mathbb{M}_{m\times d}, Q^\top Q = I_d} \sum_{j=1}^{N} \|Q^\top \mathbf{x}_j^\circ\|^2. \tag{1.7}$$

To interpret this, we return to how the matrix Q arose: its columns are an orthonormal set of d vectors in $\mathbb{R}^m$, $Q = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d]$. Thus, for any vector $\mathbf{x} \in \mathbb{R}^m$,

$$Q^\top \mathbf{x} = \begin{bmatrix} \hat{\mathbf{u}}_1 \cdot \mathbf{x} \\ \hat{\mathbf{u}}_2 \cdot \mathbf{x} \\ \vdots \\ \hat{\mathbf{u}}_d \cdot \mathbf{x} \end{bmatrix}$$

and so

$$\|Q^\top \mathbf{x}\|^2 = \sum_{k=1}^{d} (\hat{\mathbf{u}}_k \cdot \mathbf{x})^2 = \sum_{k=1}^{d} (\hat{\mathbf{u}}_k \cdot \mathbf{x})(\mathbf{x} \cdot \hat{\mathbf{u}}_k) = \sum_{k=1}^{d} \hat{\mathbf{u}}_k^\top \mathbf{x}\mathbf{x}^\top \hat{\mathbf{u}}_k = \sum_{k=1}^{d} \hat{\mathbf{u}}_k \cdot \mathbf{x}\mathbf{x}^\top \hat{\mathbf{u}}_k.$$

Combining this with (1.7), we see that we are looking to maximize (as a function of the orthonormal vectors $\{\hat{\mathbf{u}}_k\}_{k=1}^d$)

$$\sum_{j=1}^{N} \sum_{k=1}^{d} \hat{\mathbf{u}}_k \cdot \mathbf{x}_j^\circ \mathbf{x}_j^{\circ\top} \hat{\mathbf{u}}_k = \sum_{k=1}^{d} \hat{\mathbf{u}}_k \cdot \left(\sum_{j=1}^{N} \mathbf{x}_j^\circ \mathbf{x}_j^{\circ\top} \right) \hat{\mathbf{u}}_k.$$

The internal sum is a matrix we should recognize: it is (a multiple of) the **sample covariance matrix** of our data:

$$\mathbf{C} = \frac{1}{N} \sum_{k=1}^{N} (\mathbf{x}_k - \overline{\mathbf{x}}_N)(\mathbf{x}_k - \overline{\mathbf{x}}_N)^\top. \tag{1.8}$$

(Many sources use the notation $\mathbf{\Sigma}$ for the sample covariance matrix; we will avoid this notation, since it might be confusing with the standard notation Σ for the "diagonal" part in the Singular Value Decomposition of a matrix — which, we will soon see, is going to play a role here.)

REMARK 1.16. We are using the *biased* sample covariance matrix; if we instead had a factor of $\frac{1}{N-1}$, it would be an unbiased estimator of the true covariance matrix. We are not concerned about bias here, since both estimators are consistent; it will be slightly neater to use N instead of $N-1$, so we will do so consistently in the sequel.

Since a factor of $\frac{1}{N}$ does not affect where the maximum occurs, we therefore have the final form of our optimization problem:

$$Q_0 = \operatorname*{argmax}_{Q=[\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_d], Q^\top Q = I_d} \sum_{k=1}^{d} \hat{\mathbf{u}}_k \cdot \mathbf{C}\hat{\mathbf{u}}_k \tag{1.9}$$

This is a very convenient formulation! It shows that the best fit subspace is determined by the sample covariance matrix. The sample covariance matrix is a particularly nice matrix. It is a sum of rank 1 matrices each of the form $\mathbf{v}\mathbf{v}^\top$, which shows that it is symmetric. Even better: the whole matrix $\mathbf{C}$ has the form $\mathbf{C} = SS^\top$ for some (rectangular) matrix S.

LEMMA 1.17. *Let $\hat{\mathbf{X}}$ denote the $m \times N$ matrix whose columns are the centered and scaled data points $\frac{1}{\sqrt{N}}\mathbf{x}_k^\circ = \frac{1}{\sqrt{N}}(\mathbf{x}_k - \overline{\mathbf{x}}_N)$. Then $\mathbf{C} = \hat{\mathbf{X}}\hat{\mathbf{X}}^\top$.*

PROOF. We simply compute the entries of this matrix.

$$[\hat{\mathbf{X}}\hat{\mathbf{X}}^\top]_{ij} = \sum_{k=1}^{N} [\hat{\mathbf{X}}]_{ik}[\hat{\mathbf{X}}^\top]_{kj} = \sum_{k=1}^{N} [\hat{\mathbf{X}}]_{ik}[\hat{\mathbf{X}}]_{jk} = \frac{1}{N}\sum_{k=1}^{N} [\mathbf{x}_k - \overline{\mathbf{x}}_N]_i [\mathbf{x}_k - \overline{\mathbf{x}}_N]_j.$$

Comparing to (1.8) yields the result. □

COROLLARY 1.18. *The sample covariance matrix is positive semidefinite: it is symmetric and has non-negative eigenvalues.*

PROOF. The preceding lemma shows that $\mathbf{C} = \hat{\mathbf{X}}\hat{\mathbf{X}}^\top$ for some $m \times N$ matrix $\hat{\mathbf{X}}$. It is therefore symmetric: $(\hat{\mathbf{X}}\hat{\mathbf{X}}^\top)^\top = (\hat{\mathbf{X}}^\top)^\top \hat{\mathbf{X}}^\top = \hat{\mathbf{X}}\hat{\mathbf{X}}^\top$. Moroever, if λ is an eigenvalue with unit eigenvector $\hat{\mathbf{u}} \in \mathbb{R}^m$, then $\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\hat{\mathbf{u}} = \lambda\hat{\mathbf{u}}$, and so

$$\lambda = \lambda\|\hat{\mathbf{u}}\|^2 = \lambda\hat{\mathbf{u}} \cdot \hat{\mathbf{u}} = \hat{\mathbf{X}}\hat{\mathbf{X}}^\top\hat{\mathbf{u}} \cdot \hat{\mathbf{u}} = \hat{\mathbf{X}}^\top\hat{\mathbf{u}} \cdot \hat{\mathbf{X}}^\top\hat{\mathbf{u}} = \|\hat{\mathbf{X}}^\top\hat{\mathbf{u}}\|^2.$$

Thus λ is the length^2 of some vector in $\mathbb{R}^N$, and is thus ≥ 0. □

Before establishing how to completely solve optimization problem (1.9), let us now consider the special case when $d = 1$.

PROPOSITION 1.19. *Let λ_1 be the largest eigenvalue of $\mathbf{C}$. Then any unit eigenvector $\hat{\mathbf{u}}_1$ of $\mathbf{C}$ with eigenvalue λ_1 maximizes $\hat{\mathbf{u}} \mapsto \hat{\mathbf{u}} \cdot \mathbf{C}\hat{\mathbf{u}}$.*

PROOF. This is precisely the Rayleigh quotient calculation: we proved that a real symmetric matrix like $\mathbf{C}$ has at least one eigenvalue by maximizing the function $\rho(\hat{\mathbf{u}}) = \hat{\mathbf{u}} \cdot \mathbf{C}\hat{\mathbf{u}}$ over the unit sphere of length 1 vectors $\hat{\mathbf{u}}$, and showing that the resulting maximizer $\hat{\mathbf{u}}_1$ is an eigenvector of $\mathbf{C}$ with eigenvalue $\lambda_1 = \max \rho$. □

1.6. The Ky Fan Inequality

The last result in the preceding section suggests that eigenvalues and eigenvectors of the sample covariance matrix $\mathbf{C}$ should play a role in our maximization problem, and this turns out to be correct. The key is the following inequality, proved by Ky Fan in 1949, which gives a canonical upper bound for the quantity we seek to maximize.

THEOREM 1.20 (Ky Fan Inequality). *Let H be a symmetric $m \times m$ matrix, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. If $1 \leq d \leq m$, and $\{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d\}$ are orthonormal vectors in $\mathbb{R}^m$, then*

$$\sum_{k=1}^{d} \hat{\mathbf{u}}_k \cdot H\hat{\mathbf{u}}_k \leq \sum_{j=1}^{d} \lambda_j.$$

PROOF. Since H is symmetric, $\mathbb{R}^m$ has an orthonormal basis $\{\hat{\mathbf{v}}_j\}_{j=1}^m$ of eigenvectors of H; here we order them so that $H\hat{\mathbf{v}}_j = \lambda_j \hat{\mathbf{v}}_j$. We can expand the vectors $\hat{\mathbf{u}}_k$ in terms of the basis $\hat{\mathbf{v}}_j$; since the latter are orthonormal, the expansion is

$$\hat{\mathbf{u}}_k = \sum_{j=1}^{m} (\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)\hat{\mathbf{v}}_j.$$

Hence

$$H\hat{\mathbf{u}}_k = \sum_{j=1}^{m} (\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j) H\hat{\mathbf{v}}_j = \sum_{j=1}^{m} (\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)\lambda_j \hat{\mathbf{v}}_j$$

and therefore

$$\hat{\mathbf{u}}_k \cdot H\hat{\mathbf{u}}_k = \sum_{j=1}^{m} (\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)\lambda_j \hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j = \sum_{j=1}^{m} \lambda_j (\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)^2.$$

We now cleverly write $\lambda_j = \lambda_d + (\lambda_j - \lambda_d)$. Noting that (by the choice of ordering) $\lambda_j - \lambda_d \geq 0$ when $j \leq d$ and $\lambda_j - \lambda_d \leq 0$ when $j > d$, we have

$$\begin{aligned}
\hat{\mathbf{u}}_k \cdot H\hat{\mathbf{u}}_k &= \sum_{j=1}^{m} (\lambda_d + (\lambda_j - \lambda_d))(\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)^2 \\
&= \lambda_d \sum_{j=1}^{m} (\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)^2 + \sum_{j=1}^{d} (\lambda_j - \lambda_d)(\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)^2 + \sum_{j=d+1}^{m} (\lambda_j - \lambda_d)(\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)^2 \\
&\leq \lambda_d \sum_{j=1}^{m} (\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)^2 + \sum_{j=1}^{d} (\lambda_j - \lambda_d)(\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)^2. \qquad (1.10)
\end{aligned}$$

Let V denote the $m \times m$ matrix $V = [\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_m]$; since the $\hat{\mathbf{v}}_j$ form an orthonormal basis for $\mathbb{R}^m$, $V \in \mathrm{O}(m)$, so $VV^\top = I_m$. Note that

$$\sum_{j=1}^{m} (\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)^2 = \left\| \begin{bmatrix} \hat{\mathbf{v}}_1^\top \hat{\mathbf{u}}_k \\ \hat{\mathbf{v}}_2^\top \hat{\mathbf{u}}_k \\ \vdots \\ \hat{\mathbf{v}}_m^\top \hat{\mathbf{u}}_k \end{bmatrix} \right\|^2 = \left\| V^\top \hat{\mathbf{u}}_k \right\|^2 = V^\top \hat{\mathbf{u}}_k \cdot V^\top \hat{\mathbf{u}}_k = \hat{\mathbf{u}}_k \cdot VV^\top \hat{\mathbf{u}}_k = \left\| \hat{\mathbf{u}}_k \right\|^2 = 1. \quad (1.11)$$

Thus (1.10) shows that

$$\hat{\mathbf{u}}_k \cdot H\hat{\mathbf{u}}_k \leq \lambda_d + \sum_{j=1}^{d} (\lambda_j - \lambda_d)(\hat{\mathbf{u}}_k \cdot \hat{\mathbf{v}}_j)^2.$$

Therefore

$$\begin{aligned}\sum_{j=1}^{d}\lambda_j - \sum_{k=1}^{d}\hat{\mathbf{u}}_k\cdot H\hat{\mathbf{u}}_k &\geq \sum_{j=1}^{d}\lambda_j - \sum_{k=1}^{d}\left(\lambda_d + \sum_{j=1}^{d}(\lambda_j-\lambda_d)(\hat{\mathbf{u}}_k\cdot\hat{\mathbf{v}}_j)^2\right)\\ &= \sum_{j=1}^{d}(\lambda_j-\lambda_d) - \sum_{j,k=1}^{d}(\lambda_j-\lambda_d)(\hat{\mathbf{u}}_k\cdot\hat{\mathbf{v}}_j)^2\\ &= \sum_{j=1}^{d}(\lambda_j-\lambda_d)\left(1-\sum_{k=1}^{d}(\hat{\mathbf{u}}_k\cdot\hat{\mathbf{v}}_j)^2\right). \qquad (1.12)\end{aligned}$$

Finally, arguing as above: since the $\{\hat{\mathbf{u}}_k\}_{k=1}^d$ are orthonormal, they can be extended to an orthonormal basis $\{\hat{\mathbf{u}}_k\}_{k=1}^m$ of $\mathbb{R}^m$. Now arguing exactly as in (1.11) but with the roles of $\hat{\mathbf{u}}_k$ and $\hat{\mathbf{v}}_j$ reversed, we have

$$\sum_{k=1}^{d}(\hat{\mathbf{u}}_k\cdot\hat{\mathbf{v}}_j)^2 \leq \sum_{k=1}^{m}(\hat{\mathbf{u}}_k\cdot\hat{\mathbf{v}}_j)^2 = 1.$$

Thus $1-\sum_{k=1}^{d}(\hat{\mathbf{u}}_k\cdot\hat{\mathbf{v}}_j)^2 \geq 0$, and so (1.12) shows that $\sum_{j=1}^{d}\lambda_j - \sum_{k=1}^{d}\hat{\mathbf{u}}_k\cdot H\hat{\mathbf{u}}_k \geq 0$, as desired. □

REMARK 1.21. An entirely analogous proof shows that

$$\sum_{k=1}^{d}\hat{\mathbf{u}}_k\cdot H\hat{\mathbf{u}}_k \geq \sum_{j=m-d+1}^{m}\lambda_j$$

and this is the minimum.

The Ky Fan Inequality gives a nice upper bound for the quantity we're aiming to maximize; in fact, this upper bound *is* the absolute maximum, and studying the proof shows how to identify a maximizer.

COROLLARY 1.22. *Let H be a symmetric $m\times m$ matrix, with eigenvalues $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m$ and corresponding orthonormal eigenvectors $\hat{\mathbf{v}}_j$, so $H\hat{\mathbf{v}}_j=\lambda_j\hat{\mathbf{v}}_j$. For $1\leq d\leq m$,*

$$\sum_{k=1}^{d}\hat{\mathbf{v}}_k\cdot H\hat{\mathbf{v}}_k = \sum_{k=1}^{d}\lambda_k = \max_{Q=[\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_d],Q^\top Q=I_d}\sum_{k=1}^{d}\hat{\mathbf{u}}_k\cdot H\hat{\mathbf{u}}_k.$$

PROOF. The fact that the maximum is achieved on the d eigenvectors corresponding to the d largest eigenvalues can be seen from the proof of Theorem 1.20. Alternatively, and more simply, we just plug into the sum with the identities $H\hat{\mathbf{v}}_k=\lambda_k\hat{\mathbf{v}}_k$:

$$\sum_{k=1}^{d}\hat{\mathbf{v}}_k\cdot H\hat{\mathbf{v}}_k = \sum_{k=1}^{d}\hat{\mathbf{v}}_k\cdot\lambda_k\hat{\mathbf{v}}_k = \sum_{k=1}^{d}\lambda_k\|\hat{\mathbf{v}}_k\|^2 = \sum_{k=1}^{d}\lambda_k.$$

□

1.7. Principal Components

Summarizing: we have now solved the minimization problem of Definition 1.7, as follows.

THEOREM 1.23. *Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be data points in $\mathbb{R}^m$, with sample mean $\overline{\mathbf{x}}_N = \frac{1}{N}\sum_{j=1}^N \mathbf{x}_j$, and sample covariance matrix $\mathbf{C} = \frac{1}{N}\sum_{j=1}^N \mathbf{x}_j\mathbf{x}_j^\top$. Let $d \leq m$. Among all d-dimensional affine subspaces $\mathcal{A} = \boldsymbol{\mu} + V \subseteq \mathbb{R}^m$, the sum of least squares distance from the $\mathbf{x}_j$ to $\mathcal{A}$ is minimized with $\boldsymbol{\mu} = \overline{\mathbf{x}}_N$, and with V equal to the span of the d eigenvectors of the sample covariance $\mathbf{C}$ corresponding to its d largest eigenvalues. That is: if $\mathbf{C} = U\Lambda U^\top$ with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ where $\lambda_1 \geq \cdots \geq \lambda_m$, then*

$$\operatorname*{argmin}_{\substack{\boldsymbol{\mu}\in\mathbb{R}^m, \boldsymbol{\beta}_j\in\mathbb{R}^d \\ Q\in\mathbb{M}_{m\times d}, Q^\top Q = I_d}} \sum_{j=1}^N \|\mathbf{x}_j - (\boldsymbol{\mu} + Q\boldsymbol{\beta}_j)\|^2$$

is satisfied by $\boldsymbol{\mu} = \overline{\mathbf{x}}_N$, $Q = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d]$ is the $m \times d$ matrix whose columns are the first (left-most) d columns of U, and $\boldsymbol{\beta}_j = Q^\top(\mathbf{x}_j - \overline{\mathbf{x}}_N)$.

As noted: the form of the $\boldsymbol{\beta}_j$ shows that (with little surprise) the minimum is achieved by orthogonally projecting the shifted data $\mathbf{x}_j^\circ = \mathbf{x}_j - \overline{\mathbf{x}}_N$ into the given subspace. That projection is given by $QQ^\top\mathbf{x}_j^\circ = Q\boldsymbol{\beta}_j$. Thus, the vector $\boldsymbol{\beta}_j = Q^\top\mathbf{x}_j^\circ$ is the vector in $\mathbb{R}^d$ whose components are the coordinates of the best d-dimensional approximation of $\mathbf{x}_j^\circ$, in the basis $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d$.

DEFINITION 1.24. *The eigenvectors $\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_m$ of $\mathbf{C}$ are call the* **principal components** *or* **principal component axes** *of the data; as usual, they are ordered by decreasing eigenvalues. Thus, the d-dimensional subspace which best (in least squares sense) approximates the (centered) data is the span of the first d principal components. We call the subspace*

$$\mathcal{P}_d := \operatorname{span}\{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d\}$$

the **rank-d principal space** *for the data.*

Let $Q_d = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d]$; then $\mathcal{P}_d = \operatorname{Col}(Q_d)$. The orthogonal projection onto $\operatorname{Col}(Q_d)$ is $Q_dQ_d^\top$. For any data point $\mathbf{x}_j$, we have

$$Q_dQ_d^\top\mathbf{x}_j = Q_dQ_d^\top(\overline{\mathbf{x}}_N + \mathbf{x}_j^\circ) = Q_d(\overline{\boldsymbol{\beta}}_N^d + \boldsymbol{\beta}_j^d)$$

where $\overline{\boldsymbol{\beta}}_N^d = Q_d^\top\overline{\mathbf{x}}_N$ and $\boldsymbol{\beta}_j^d = Q_d^\top\mathbf{x}_j^\circ$ are in $\mathbb{R}^d$. We call these vectors the **rank-d principal mean** *and* **rank-d principal coordinates** *of the data. That is: $\boldsymbol{\beta}_j^d$ gives the coefficients of the projection of $\mathbf{x}_j^\circ$ in $\mathcal{P}_d$, in terms of the basis $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d$, while $\overline{\boldsymbol{\beta}}_N^d$ gives the affine shift of the projected data.*

REMARK 1.25. To be totally clear: the coordinate vector $\boldsymbol{\beta}_j^d$ is given by

$$\boldsymbol{\beta}_j^d = \begin{bmatrix} \hat{\mathbf{u}}_1 \cdot \mathbf{x}_j^\circ \\ \hat{\mathbf{u}}_2 \cdot \mathbf{x}_j^\circ \\ \vdots \\ \hat{\mathbf{u}}_d \cdot \mathbf{x}_j^\circ \end{bmatrix}$$

and so indeed

$$Q_d\boldsymbol{\beta}_j^d = \sum_{k=1}^d (\hat{\mathbf{u}}_k \cdot \mathbf{x}_j^\circ)\hat{\mathbf{u}}_k$$

is the orthogonal projection of $\mathbf{x}_j^\circ$ onto $\mathcal{P}_d = \operatorname{span}\{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d\} = \operatorname{Col}(Q_d)$.

Why call the vectors $\hat{\mathbf{u}}_j$ "principal components" instead of just "eigenvectors"? The reason is that they arise in a slightly different (but equivalent) statistical context. There is a different

optimization problem one might want to do with the data. Instead of looking for a "best fit" subspace (of given dimension), we instead look for a subspace (of given dimension) that maximizes the sample variance of the orthogonally projected data. (We have omitted the word "affine" in this context, since translating the data will not affect the variance.) First, to be clear on definitions:

DEFINITION 1.26. *Given a random vector* $\mathbf{Y} \in \mathbb{R}^d$*, the* **variance** *of* $\mathbf{Y}$ *is*

$$\mathrm{Var}(\mathbf{Y}) = \mathbb{E}\left(\|\mathbf{Y} - \mathbb{E}(\mathbf{Y})\|^2\right).$$

Equivalently, if $\mathbf{Y}$ *has covariance matrix* C*, then* $\mathrm{Var}(\mathbf{Y}) = \mathrm{Tr}(C)$.

If $\{\mathbf{y}_j\}_{j=1}^N$ *are data points in* $\mathbb{R}^d$*, their (biased)* **sample variance** $S_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ *is*

$$S_N(\mathbf{y}_1, \ldots, \mathbf{y}_N) = \frac{1}{N}\sum_{j=1}^N \|\mathbf{y}_j - \overline{\mathbf{y}}_N\|^2.$$

Equivalently, if the sample covariance matrix of the data is $\mathbf{C}$*, the sample variance is* $\mathrm{Tr}(\mathbf{C})$.

REMARK 1.27. To justify the last statements, that the variance is the trace of the covariance matrix: the sample covariance matrix is

$$\mathbf{C} = \hat{\mathbf{Y}}\hat{\mathbf{Y}}^\top$$

where $\hat{\mathbf{Y}}$ has columns $\frac{1}{\sqrt{N}}\mathbf{y}_j^\circ = \frac{1}{\sqrt{N}}(\mathbf{y}_j - \overline{\mathbf{y}}_N)$; in other words, $[\hat{\mathbf{Y}}]_{ij} = \frac{1}{\sqrt{N}}[\mathbf{y}_j^\circ]_i$ is the ith component of the vector $\frac{1}{\sqrt{N}}\mathbf{y}_j^\circ$. The diagonal entries of this matrix are

$$[\mathbf{C}]_{ii} = \sum_{j=1}^N [\hat{\mathbf{Y}}]_{ij}[\hat{\mathbf{Y}}^\top]_{ji} = \sum_{j=1}^N [\hat{\mathbf{Y}}]_{ij}^2$$

and so

$$\mathrm{Tr}(\mathbf{C}) = \sum_{i=1}^d [\mathbf{C}]_{ii} = \sum_{i=1}^d\sum_{j=1}^N [\hat{\mathbf{Y}}]_{ij}^2 = \sum_{j=1}^N\sum_{i=1}^d [\hat{\mathbf{Y}}]_{ij}^2 = \sum_{j=1}^N\sum_{i=1}^d \frac{1}{N}[\mathbf{y}_j^\circ]_i^2 = \frac{1}{N}\sum_{j=1}^N \|\mathbf{y}_j^\circ\|^2.$$

Given an orthonormal set of d vectors in $\mathbb{R}^m$, as usual put together in a matrix $Q \in \mathbb{M}_{m\times d}$ satisfying $Q^\top Q = I_d$, we can orthogonally project our data $\mathbf{x}_j$ into the subspace spanned by these vectors, giving us projected data $\mathbf{y}_j = QQ^\top\mathbf{x}_j$. We can then look for the subspace of dimension d (given by orthonormal basis matrix Q) which maximizes the (sample) variance of the projected data. The solution turns out to be exactly the same as the solution to the least squares best fit problem above.

PROPOSITION 1.28. *Let* $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ *be data points in* $\mathbb{R}^m$*, and let* $d \leq m$*. The* d*-dimensional subspace of* $\mathbb{R}^m$ *that maximizes the variance of the projected data is precisely the rank-*d *principal space* $\mathcal{P}_d$ *of the data.*

PROOF. The problem here is

$$\operatorname*{argmax}_{Q\in\mathbb{M}_{m\times d}, Q^\top Q = I_d} S_N(QQ^\top\mathbf{x}_1, \ldots, QQ^\top\mathbf{x}_N).$$

Note that if $\mathbf{y}_j = QQ^\top\mathbf{x}_j$ then

$$\overline{\mathbf{y}}_N = \frac{1}{N}\sum_{j=1}^N \mathbf{y}_j = \frac{1}{N}\sum_{j=1}^N QQ^\top\mathbf{x}_j = QQ^\top\left(\frac{1}{N}\sum_{j=1}^N \mathbf{x}_j\right) = QQ^\top\overline{\mathbf{x}}_N.$$

Thus

$$\|\mathbf{y}_j - \overline{\mathbf{y}}_N\|^2 = \|QQ^\top \mathbf{x}_j - QQ^\top \overline{\mathbf{x}}_N\|^2 = \|QQ^\top \mathbf{x}_j^\circ\|^2.$$

Thus, we are looking for

$$\operatorname*{argmax}_{Q \in \mathbb{M}_{m\times d}, Q^\top Q = I_d} \frac{1}{N} \sum_{j=1}^{N} \|QQ^\top \mathbf{x}_j^\circ\|^2.$$

Comparing this to (1.7) (and the calculation directly above that equation) shows that the solution is the same as the one for the least squares problem, as claimed. □

What is slightly more interesting about this formulation is that the maximum *value* takes on a new meaning.

COROLLARY 1.29. *The maximum sample variance of any projection of the data into a d-dimensional subspace is*

$$\max_{Q \in \mathbb{M}_{m\times d}, Q^\top Q = I_d} S_N(QQ^\top \mathbf{x}_1, \ldots, QQ^\top \mathbf{x}_N) = \lambda_1 + \cdots + \lambda_d$$

where $\lambda_1, \ldots, \lambda_d$ are the d largest eigenvalues of the sample covariance matrix of the data.

PROOF. This is precisely the statement of Corollary 1.22 in this context. □

This gives us some idea of what the eigenvalues of the positive semidefinite matrix $\mathbf{C}$ here represent: they quantify the amount of the sample variance contained in the different subspaces. Indeed, the sample variance of the full data set is $\operatorname{Tr}(\mathbf{C})$ (see Definition 1.26), and this is equal to the sum of all of the eigenvalues of $\mathbf{X}$, $\lambda_1 + \lambda_2 + \cdots + \lambda_m$. We therefore interpret the above as saying that the first principal component $\hat{\mathbf{u}}_1$ is responsible for λ_1 variance in the data, the second principal component $\hat{\mathbf{u}}_2$ is responsible for λ_2 variance in the data, and so forth. Alternatively, we might say that the *proportion of the variance* that principal component $\hat{\mathbf{u}}_k$ is responsible for is

$$\text{Proportion of variance in } \operatorname{span}\{\hat{\mathbf{u}}_k\} = \frac{\lambda_k}{\sum_{j=1}^m \lambda_j} = \frac{\lambda_k}{\operatorname{Tr}(\mathbf{C})}.$$

REMARK 1.30. The principal components, being eigenvectors of a symmetric matrix, are orthogonal. In terms of covariance, this means that the projections of the data into the spans of different principal components are *uncorrelated*. It is for this reason that the variances along different principle components add up to the variance attributed to the span of them.

1.8. The Singular Value Decomposition

The principal components are eigenvectors of the sample covariance matrix of the data. Recall, from Lemma 1.17, that the sample covariance matrix can be computed as $\hat{\mathbf{X}}\hat{\mathbf{X}}^\top$, where $\hat{\mathbf{X}}$ is the $m \times N$ matrix whose columns are the centered data points scaled by $\frac{1}{\sqrt{N}}$. That is: given data $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in $\mathbb{R}^m$, let

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{M}_{m\times N}.$$

This matrix is sometimes called the **feature matrix** or (less appropriately) **feature vector** in statistics books. The entires of each data point $\mathbf{x}_j$ are the *features* that are being measured. The matrix $\hat{\mathbf{X}}$ is computed from $\mathbf{X}$ by subtracting off the column $\overline{\mathbf{x}}_N$ from each column of the matrix and then scaling by $\frac{1}{\sqrt{N}}$; that is, we subtract the rank-1 matrix all of whose columns are $\overline{\mathbf{x}}_N$. This can be written in terms of the vector $\mathbb{1} \in \mathbb{R}^N$ of all ones:

$$\hat{\mathbf{X}} = \frac{1}{\sqrt{N}}(\mathbf{X} - \overline{\mathbf{x}}_N \mathbb{1}^\top).$$

Now, we may compute the singular value decomposition (SVD) of $\hat{\mathbf{X}}$:

$$\hat{\mathbf{X}} = U\Sigma V^\top$$

where $U \in \mathrm{O}(m)$, $V \in \mathrm{O}(N)$, and $\Sigma \in \mathbb{M}_{m\times N}$ is "diagonal" (more properly, the top square $N \times N$ part of Σ is diagonal) with $[\Sigma]_{ii} = \sigma_i$, the singular values of $\hat{\mathbf{X}}$, canonically ordered non-increasingly $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N \geq 0$. Then

$$\mathbf{C} = \hat{\mathbf{X}}\hat{\mathbf{X}}^\top = (U\Sigma V^\top)(V\Sigma^\top U^\top) = U\Lambda U^\top$$

where $\Lambda = \Sigma\Sigma^\top$ is the $m \times m$ diagonal matrix whose entries are $[\Lambda]_{ii} = \sigma_i^2$ for $i \leq N$ and 0 for $N < i \leq m$. This means we can immediately recast Definition 1.24, Proposition 1.28, and Corollary 1.29 as follows.

PROPOSITION 1.31. *The principal components are the* left singular vectors $\hat{\mathbf{u}}_k$ *of* $\hat{\mathbf{X}}$. *If* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N$ *are the singular values of* $\hat{\mathbf{X}}$, $\operatorname{span}\{\hat{\mathbf{u}}_k\}$ *is responsible for* σ_k^2 *variance in the data (or proportion* $\sigma_k^2/(\sigma_1^2 + \cdots + \sigma_N^2)$ *of the total).*

What a lovely coincidence! We are accustomed to using σ to denote both singular values and standard deviations; in PCA/SVD, the notational collision works out perfectly.

REMARK 1.32. Some authors actually define the principal components to be $\sigma_k\hat{\mathbf{u}}_k$; that is, they are the nonzero columns of $U\Sigma$. This has some geometric meaning in terms of the SVD: there is a decomposition of $\hat{\mathbf{X}}$ as a product $\hat{\mathbf{X}} = PQ$ where $Q \in \mathrm{O}(N)$ and $P \in \mathbb{M}_{m\times N}$ has orthogonal columns with decreasing lengths; namely, $P = U\Sigma$ and $Q = V^\top$. From a statistical standpoint, this just means weighting each principal component by the standard deviation it imbues the data; thus, in this model, all principal components carry equal proportions of the variance. We refer to these as *standardized principal components*; we will largely work with the ordinary principal components of Definition 1.24.

Aside from notation, one reason to use the SVD, rather than simply diagonalizing $\mathbf{C}$, is computational complexity. Either way: to begin we must compute the matrix $\hat{\mathbf{X}}$, which means computing $\overline{\mathbf{x}}_N$ ($O(mN)$ flops), then subtract it from each entry of $\mathbf{X}$ and divide each by $\sqrt{N}$ ($O(mN)$ flops). Now, if we wish to diagonalize $\mathbf{C}$ we must compute it: $\mathbf{C} = \hat{\mathbf{X}}\hat{\mathbf{X}}^\top$ requires $O(m^2N)$ flops, totalling $O(m^2N)+O(mN) = O(m^2N)$. Recall that $m \gg N$. On the other hand, if we want to find the SVD, it's just as good to compute $\hat{\mathbf{X}}^\top\hat{\mathbf{X}}$, which only requires $O(N^2m)$ flops, *much cheaper*. Now we must diagonalize this matrix. There are a number of numerical ways to do this to whatever desired accuracy: the QR algorithm; tridiagonalization (via Householder); the power method; and others. All of these compute the eignevalues σ_j^2 and eigenvectors $\hat{\mathbf{v}}_j$ to desired accuracy at the cost of $O(N^3)$ flops. Once this is accomplished, the desired left eigenvectors $\hat{\mathbf{u}}_j$ are simply $\hat{\mathbf{u}}_j = \frac{1}{\sigma_j}\hat{\mathbf{X}}\hat{\mathbf{v}}_j$, each of which costs $O(mN)$ flops to compute. Since $N < m$, there are only at most N non-zero singular values; the remaining $\hat{\mathbf{u}}_j$ are all eigenvectors of $\mathbf{C}$ with eigenvalue 0, and we do not care about these (as those principal components carry 0 variance). Hence, we've computed all the relevant principal components at a total cost of $O(m^2N) + O(N^3) + O(mN) = O(mN^2)$ — much less than the cost $O(m^2N)$ of even computing the matrix $\mathbf{C}$!

Even so: we typically do not need *all* the principal components. We will be looking for only the d top principal components, where $d \ll N$. So at the last step, computing only d of the $\hat{\mathbf{u}}_j$ from the $\hat{\mathbf{v}}_j$ already computed only requires $O(dmN)$ flops. From the above steps, though, it already took $O(mN^2)$ flops to compute the σ_j and $\hat{\mathbf{v}}_j$, and $O(mN^2)$ dominates $O(dmN)$. Nevertheless, there are much more clever ways of doing *truncated* SVD that compute (to desired accuracy)

the d top eigenvectors without computing any further ones; hence, the entire computation of the top d principal components can be accomplished in only $O(dmN)$ flops using truncated SVD algorithms. Even more surprising: by using a *randomized* algorithm (only selecting a small number of columns from $\hat{\mathbf{X}}$ randomly to work with), it is possible to compute, to desired accuracy, the top d principal components in $O(mN \log d + (m+N)d^2)$ flops. (This is very recent work, published in 2009, by Halko, Martinsson, and Tropp.)

1.9. Dimension Reduction and Visualization

Returning to our original problem: we have a high dimensional data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in $\mathbb{R}^m$ which we would like to analyze for structure. We can now look at projections of the data into the best fit (or equivalently maximal variance preserving) low dimensional affine subspaces. Choosing the desired dimension d for the projection is a very subtle problem which we will tackle after the next chapter. Once we have made this choice, we compute the top d principal components $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d$, and project the (centered) data into the subspace they span. Of course, we have now just collapsed the data onto some low-rank subspace still in the high-dimensional ambient space $\mathbb{R}^m$; how does this help?

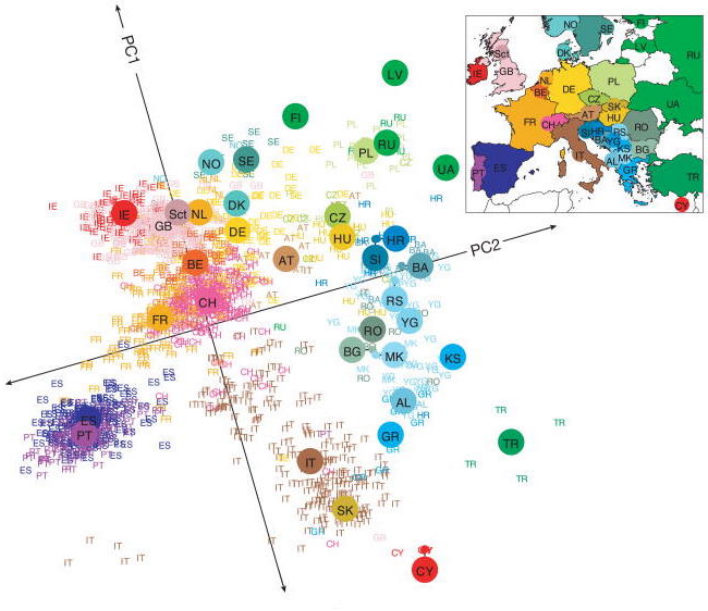
If we have a d-dimensional space, it is isomorphic to $\mathbb{R}^d$, so we can choose any isomorphism and use it to represent the (projected, centered) data in $\mathbb{R}^d$ instead. An isomorphism of vector spaces is equivalent to a choice of bases in each one, and we have canonical bases: the subspace was chosen to be spanned by the orthonormal principal components, so we use them as a basis there, and of course we use the standard basis in $\mathbb{R}^d$. Referring to Definition 1.24, letting $Q_d = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_d]$, the transformation just described is

$$\mathbb{R}^m \ni \mathbf{x}_j \mapsto \mathbf{x}_j^\circ \mapsto Q_d^\top \mathbf{x}_j^\circ = \boldsymbol{\beta}_j \in \mathbb{R}^d.$$

That is: we represent the projected data via their *rank-d principal coordinates*. Better yet: if we reincorporate the mean of the data, we should plot $Q_d^\top \mathbf{x}_j = Q_d^\top \overline{\mathbf{x}}_N + Q_d^\top \mathbf{x}_j^\circ = \overline{\boldsymbol{\beta}}_N^d + \boldsymbol{\beta}_j^d$: the *rank-$d$ principal coordinates shifted by the rank-d principal mean*.

EXAMPLE 1.33. In the 2008 Nature paper *Genes mirror geography within Europe*, the authors studied a large, high-dimensional dataset of genetic markers from people with European grandparents. Each of the $N = 3,192$ individuals were interviewed about where their grandparents had lived precisely; this information was recorded but set aside. Then genetic samples were taken from the study participants, and were genotyped for $m > 500,000$ genetic markers. This means that tests were conducted on m known sites in their genomes to see which amino acid (from two possible choices) was present in their DNA at that site. In fact, each person has two copies of each chromosome (from their two parents), and so the data the study recorded was whether, at each site, the two markers were both of amino acid type 1, both of type 2, or one of each. These features were recorded as a 0, 1, or 2. Thus, the recored data points were $\mathbf{x}_j \in \{0,1,2\}^m$ for $1 \leq j \leq N$.

The authors performed PCA on this data (using a widely available software package called `smartpca`), choosing the target dimension to be $d = 2$ (which is the most common choice). They then plotted the N principal coordinates $\boldsymbol{\beta}_j^2 = [\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2]^\top \mathbf{x}_j^\circ$ in $\mathbb{R}^2$. Each of these N points corresponds to one of the original N subjects in the study; they finally colored the points according to the country of origin of each subject's grandparents. The figure below is the resulting visualization of the data, which bears a striking resemblance to a geographic map of Europe.



The authors decided to rotate and reflect the data: that is, technically, they plotted the points $R\boldsymbol{\beta}_j^2$ where R is reflection across the the second axis and then ccw rotation by about $107°$. In the figure, they plotted the transformed axes, which they have labeled PC1 and PC2. They did this in order to make the data match as closely as possible to the map of Europe (ad hoc, by superimposing).

This analysis is very suggestive that the rank-2 principal coefficients of the data correlate strongly with *lattitude* (or, more precisely, distance in the direction $17°$ NNW) and *longitude* (or, more precisely, distance in the direction $17°$ ENE).

This appears to be *magic*: there is a map of Europe encoded in their genes! A little thought will show that this is not magic, and makes good sense. Genetic variation between people is quite high; in a large and heterogeneous population, this variation between individuals appears to be mostly random noise. In European human populations of the early 20th Century and earlier, people did not move around very much. Local populations therefore become more genetically homogeneous. Thus, what PCA is telling us is that the most prominent non-random factors determining genetic drift are geographic separation: (roughly) north-south accounting for the most variation, and (roughly) east-west for the next largest component in variation. This suggests that what we are seeing here is the remnant of an ancient migration, and the data suggests this migration went in a SSE to NNW direction. This is consistent with anthropological studies showing just such migration occurred, approximately 15,000 years ago, at the end of the last ice age.

An interesting follow-up would be to look at the third principal component, and try to identify some additional variable that it measures. There might be no third feature: the third principal component may be consumed by the noise. If there is another feature, it must be one uncorrelated with latitude and longitude in Europe. Altitude? Proximity to water? Abundance of certain food staples? Maybe there's another Nature paper waiting to be written.

REMARK 1.34. It is tempting to interpret the above interpretation as a *predictor* of geographic origin. (Indeed, the authors present it this way, stating that more than half of the points landed within 400km of the hometown of the corresponding subject's grandparents. The caveat is that this only worked if all four grandparents lived in the same place; if their two sets of grandparents lived in different places, the principal coefficient tended to appear somewhere in between them on the

map.) One might dream of being able to test the same genetic markers in the blood of a random person of European descent, and from them tell her, within a few kilometers, where her ancestors lived in Europe. One subtlety is that the map is not encoded in the genes of *any one person*; it is only encoded in the feature data from the full population. (That's the part that really seems like magic.) Still, one could add the new random person's genetic markers as a new data point and calculate its principal coefficients (using the principal components from the previous data set); or one could add the point to the original dataset and redo the PCA from the start (since it is only one point added to $N > 3,000$, unless it is *wildly* different from the others, we would not expect the principal components to change much). This is an interesting idea which could lead to more research projects.

1.10. Choosing the Appropriate Dimension d

Let's summarize this chapter. We have some high dimensional data $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ in $\mathbb{R}^m$ (in our paradigm, $m \gg N$, although none of the above calculations require this). We have reason to believe that there are deterministic points ("signal") $\mathbf{t}_1, \dots, \mathbf{t}_N \in \mathbb{R}^m$ that all lie in some affine subspace $\mathcal{A}_d$ of dimension $d \ll m$ such that the recorded data $\mathbf{x}_j$ are in fact samples of random variables

$$\mathbf{X}_j = \mathbf{t}_j + \mathbf{Z}_j \tag{1.13}$$

where $\mathbf{Z}_j$ are independent random variables ("noise") of mean 0. (The mean 0 assumption is mostly for convenience; if $\mathbb{E}(\mathbf{Z}_j) = \boldsymbol{\mu}_j$ then we would instead use the centered noise random variable $\mathbf{Z}_j^\circ = \mathbf{Z}_j - \boldsymbol{\mu}_j$, and incorporate $\mathbf{t}_j \mapsto \mathbf{t}_j + \boldsymbol{\mu}_j$ as part of the "signal".) We have seen (Midterm 1 makes this explicit) that, under the assumption of i.i.d. Gaussian noise $\mathbf{Z}_j$ (which is sometimes called *white noise*), the MLE for the unknown d-dimensional affine subspace $\mathcal{A}_d$ is found through **Principal Component Analysis**: $\mathcal{A}_d = \overline{\mathbf{x}}_N + \mathrm{Col}(Q_d)$ where $Q_d = [\hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_d]$ are the first d **principal components**: the d eigenvectors of the sample covariance matrix of the data with the d largest eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_d^2$. It is then a matter of (efficient) calculation to find (good approximations of) these principal component vectors, and then project the data into this affine subspace to view it (via its **principal coordinates** $Q_d^\top \mathbf{x}_j$) in $\mathbb{R}^d$.

What this method cannot do, a priori, is determine the best d. Indeed, one can always take $d = m$ to get a *perfect match* for the data, which defeats the purpose. How, then can we discern from the data what a good choice for what is the real number of degrees of freedom, i.e. the real number of *statistically significant factors* in the data?

The answer lies in Proposition 1.28 and the following discussion: the subspace $\mathcal{A}_d$ (or any translate of it, in particular $\mathrm{Col}(Q_d)$) is the d-dimensional subspace that maximizes the variance of the (projected) data among all d-dimensional subspaces. More precisely: the eigenvalue σ_k^2 oft he sample covariance matrix is precisely equal to the sample variance of the projected data $\{P_k \mathbf{x}_j\}_{j=1}^N$, where P_k is the orthogonal projection onto $\mathrm{span}\{\hat{\mathbf{u}}_k\}$, the kth principal component. This is relevant to our question because of the model (1.13). One might think that *variance* is associated only to the random part $\mathbf{Z}_j$; but here we are talking about sample variance, and after all, we are not suggesting all the $\mathbf{t}_j$ are equal! In fact, even if $\mathbf{Z}_j = 0$, the deterministic data $\mathbf{t}_1, \dots, \mathbf{t}_N$ likely carries a lot of sample variance, indicating that it varies among the indices. (To make the point clear, many authors use the term *variation* instead of *variance* here.) In our model, there is noise as well, but we will assume that the (random) variance of the noise is much smaller than the variation of the signal (in other words: we posit a **high signal-to-noise ratio**. Under that assumption, it is natural to expect that the directions with highest variance are related to the signal, rather than the noise.

The quantitative question now becomes: *where do we draw the line between signal and noise*? We expect that the principal components, the directions of highest variation, are in the subspace where the *true* data $\mathbf{t}_j$ live; but at some point, as the variances σ_k^2 decrease, the noise overtakes the signal. Our job is to find the largest d for which σ_d^2 is still due to variation in the true signal; for $k > d$, the variance σ_k^2 is the result of random noise. How do we accomplish this?

In real world data sets, most sources simply take $d \in \{1, 2, 3\}$; not necessarily because the data suggests the "real" dimension is ≤ 3, but because these are the only dimensions in which we can visually present the data. Even if we place this constraint, we still want to know what the "right" d is; after all, we may choose to plot PC1 and PC2, but it might turn out that the data is really only 1-dimensional, and PC2 is random noise, in which case our plot will be misleading and confusing, hunting for structure (in the second dimension) where there is none.

We therefore hope to find some extrinsic feature in the behavior of the σ_k^2 that suggests a shift in behavior as k increases. The simplest approach is to plot the values σ_k^2 as a function of $k \in \{1, \ldots, N\}$ (or perhaps with a much smaller range) to look for some kind of shift. This approach was first proposed by British psychologist Raymond Cattell in 1966. Cattell was a pioneer in the use of advanced statistical methods in psychology, and a father of what is now called "factor analysis" (essentially the term used in psychology for PCA). The (piecewise-linearized) plot of $k \mapsto \sigma_k^2$ is known as a **scree plot**. (A *scree* is a sloping mass of loose rocks at the bottom of a cliff, which comes from the Old Norse term for a landslide.)

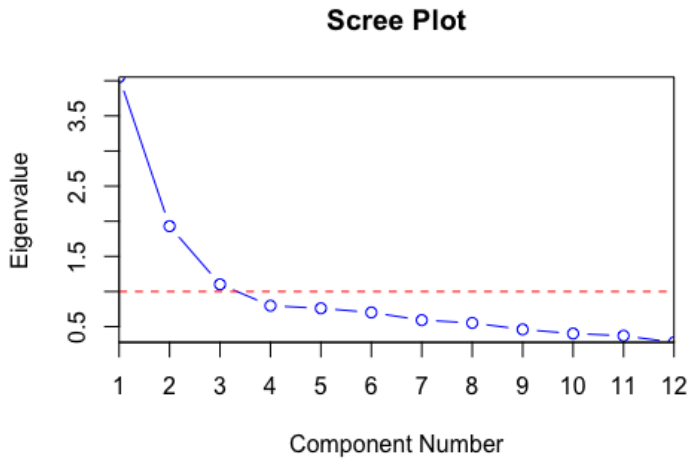


FIGURE 1.1. An example of a scree plot for the 12 largest eigenvalues, produced in R.

The idea, proposed by Cattell, is to look for an "elbow" in the scree plot: a point where the decline in σ_k^2 levels off and becomes less steep. This point of inflection is supposed to indicate the shift from variation due to structure to variance due to random noise. This methodology is problematic at best; even if it were robustly indicative of something, it is highly subjective to decide where the "elbow" is in most cases, leading to researchers (consciously or unconsciously) choosing the cut-off in line with their biases from the conclusions they *wish to draw*. For this reason, I would class the "elbow" rule as mostly pseudoscience.

REMARK 1.35. There is a deeper reason the elbow rule should be entirely tossed out. As we will see in the next chapter, data that is *purely random* produces scree plots that always have a noticeable "elbow"; hence, this "elbow" is not indicative of a change from structure to randomness.

To avoid some of these problems, a number of more systematic procedures for determining a cutoff have been proposed over the years. The simplest is the *Kaiser criterion*, which posits that the cut-off occurs when $\sigma_k^2 < 1$. (The idea here is that the *empirical average eigenvalue is* 1 (we will see this fact early in the next chapter), and so the statistically meaningful PC's are the ones whose variation is above average. (The dotted line in Figure 1.1 indicated the Kaiser criterion, which in this case coincides with a unique visible "elbow".) This method often produces too many factors: many of the $\sigma_k^2 > 1$ may still be the result of random noise. Some more robust variations include computing confidence intervals for each σ_k and placing the cutoff at the largest k for which the entire confidence interval about σ_k^2 is > 1; this is still not very robust.

The truth is that scree plots simply do not convey the right information to make a meaningful judgment of the appropriate d. As we will see in the next chapter, a much better method is to plot a *histogram* of the eigenvalues (or singular values) of the sample covariance matrix; from this representation, a robust (and visually striking) separation between *noise* and *signal* can usually be seen. The end goal of these notes is to explain this phenomenon in detail. The first important task is to *understand what noise without signal looks like*; after all, that will give us a baseline to compare a real data set to. This brings us to the next chapter, on fully **random matrices**.

CHAPTER 2

Random Matrices

We wish to investigate the behavior of the eigenvalues of large (i.e. high dimensional) random matrices. The specific models of interest to us, as motivated in Section 1.10 of the previous chapter, are sample covariance matrices $\hat{\mathbf{X}}^\top \hat{\mathbf{X}}$ where the columns of the $m \times N$ matrix $\mathbf{X}$ are i.i.d. standard normal random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ (and $\hat{\mathbf{X}} = \frac{1}{\sqrt{N}}(\mathbf{X} - \overline{\mathbf{Z}}_N \mathbb{1}^\top)$. Let us first dispense with the sample mean correction: if $\mathbf{Z}_j \sim \mathscr{N}(\mathbf{0}, I_N)$ are all independent, then by the strong law of large numbers,

$$\overline{\mathbf{Z}}_N = \frac{1}{N} \sum_{j=1}^{N} \mathbf{Z}_j \to \mathbf{0}.$$

Since we are interested in large-N behavior, we will therefore lose nothing by eliminating this term, and instead simply use the matrix $\hat{\mathbf{X}} = \frac{1}{\sqrt{N}}\mathbf{X}$.

A few calculations show that, with $m \geq N$ large, calculating the eigenvalues exactly is a completely hopeless task.

EXAMPLE 2.1. Consider the case $m = N = 2$. Then

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$$

where the four entries X_{ij} are all independent $\mathscr{N}(0,1)$ random variables. Thus

$$\hat{\mathbf{X}}^\top \hat{\mathbf{X}} = \frac{1}{2}\mathbf{X}^\top \mathbf{X} = \frac{1}{2} \begin{bmatrix} X_{11}^2 + X_{21}^2 & X_{11}X_{12} + X_{21}X_{22} \\ X_{11}X_{12} + X_{21}X_{22} & X_{12}^2 + X_{22}^2 \end{bmatrix}. \tag{2.1}$$

The diagonal entries are sums of squares of independent standard normal random variables, hence they are χ^2 random variables (with 2 degrees of freedom); and they are independent from each other. But the (equal) off diagonal entries are somewhat more complicated. A bit of work (completing the square) will show they are also χ^2 random variables (this time with 4 degrees of freedom). But they are manifestly not independent from the diagonal entries; there are lots of correlations.

It is possible to completely describe the joint distribution of the entries, but as you can see it is already not a simple task. But that's just the distribution of the *entries*; we are interested in the eigenvalues! The matrix above is symmetric, so it has the form

$$\begin{bmatrix} 2a & c \\ c & 2b \end{bmatrix}$$

for real numbers a, b, c (where the factors of 2 have been included to make the outcome of the following calculation slightly simpler). The characteristic polynomial of this generic matrix is $(\lambda - a)(\lambda - b) - c^2 = \lambda^2 - (a+b)\lambda + (ab - c^2)$, and so the eigenvalues are

$$\lambda_\pm = a + b \pm \sqrt{(a-b)^2 + c^2}.$$

Substituting in the entries from (2.1), we therefore have

$$4\lambda_\pm = X_{11}^2 + X_{12}^2 + X_{21}^2 + X_{22}^2 \pm \sqrt{(X_{11}^2 + X_{21}^2 - X_{12}^2 - X_{22}^2)^2 + 4(X_{11}X_{12} + X_{21}X_{22})^2}.$$

That's quite a doozy of an expression. (One might hope that expanding out the quartic polynomial under the square root would result in some nice cancelations or simplifications; alas, nothing really helpful occurs.) So: given four independent $\mathcal{N}(0,1)$ random variables X_{ij}, what is the joint distribution of the two random variables $\lambda_\pm$ above? One might imagine, with a great deal of sweat, it might be possible to grind out a precise description of the joint density of these two random variables. It's not clear what we might learn from the resultant over-complicated expression. And this is just the extremely low dimensional case $m = N = 2$! Our goal is to understand what happens for $m \geq N \gg 2$; quickly, even the pretense of exact calculations will be gone, since when $N > 5$ there are no formulas possible for the roots of the characteristic polynomial.

As the above example shows, we cannot hope to understand the behavior of the eigenvalues by calculating their joint distribution directly. (That's not entirely true; we will see later that there are cases where the spectral theorem can be used to great effect. But that is still not the "right" way to proceed.) The eigenvalues are highly non-linear functions of the entries. But it is not our goal to understand their joint distribution; what we want is to understand their aggregate statistical behavior. We want to understand the *histogram* of all the eigenvalues together.

2.1. Histograms and Linear Statistics

We want to understand the "distribution" of a collection of points $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_N\}$ in the real line. This is not quite "distribution" in the probability sense (although it is related): the points in question need not be random (although for us they will be: the eigenvalues of a random matrix). Rather, we mean some way of understanding the overall behavior of the points as a group: where they crowd together, where they pull apart, where they are absent. A great tool for summarizing this information is a *histogram*.

DEFINITION 2.2. *Let $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_N\}$ be a finite collection of points in $\mathbb{R}$, canonically labeled in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. Let $[a, b]$ be a closed interval containing all the points in $\boldsymbol{\lambda}$, and let $\Pi = \{a = t_0 < t_1 < \cdots < t_p = b\}$ be a partition of $[a, b]$. The (normalized)* **histogram** *$h_{\Pi,\boldsymbol{\lambda}}$ of the points $\boldsymbol{\lambda}$, relative to the partition Π, is the function defined by*

$$h_{\Pi,\boldsymbol{\lambda}}(t) = \frac{1}{N}\#\{j\colon t \text{ and } \lambda_j \text{ are in the same partition interval of } \Pi\}.$$

A histogram $h_{\Pi,\boldsymbol{\lambda}}$ is a step function: it is constant on each partition interval $(t_{i-1}, t_i]$ in the fixed partition Π. These partition intervals are usually called **bins**, and are often (but not always) chosen to be all of equal length, in which case the partition is determined by the base interval $[a, b]$ (which is often taken to be $[\lambda_N, \lambda_1]$, or possibly $[\lfloor\lambda_N\rfloor, \lceil\lambda_1\rceil]$) and the number of bins. For example: given a column vector of real number $\boldsymbol{\lambda}$, the MATLAB command `hist(`$\boldsymbol{\lambda}$`,50)` will produce the histogram of $\boldsymbol{\lambda}$ with the equal-length partition containing 50 bins, and base interval $[\lfloor\lambda_N\rfloor, \lceil\lambda_1\rceil]$.

REMARK 2.3. It is more customary for histograms to take values in the non-negative integers; we have chosen to normalize ours so they count the *proportion* of points in each bin, rather than the number.

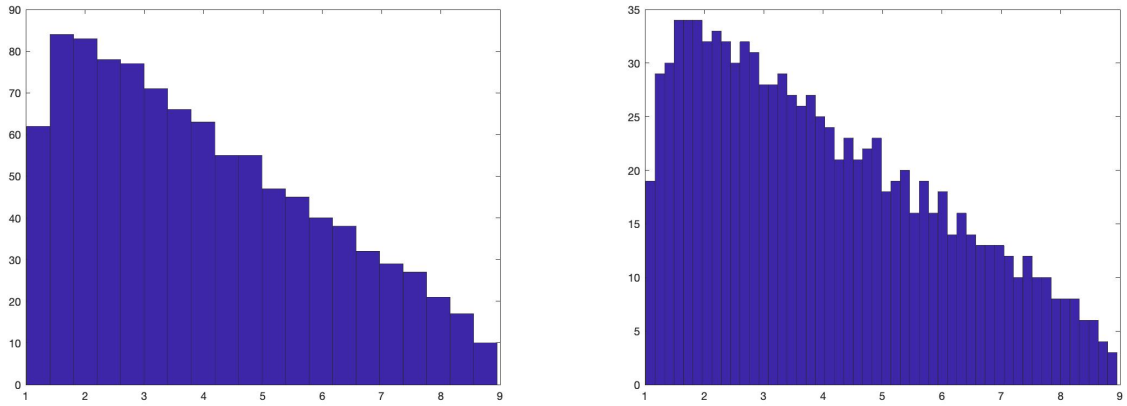


FIGURE 2.1. The graphs of two (un-normalized) histograms, with 20 bins and 50 bins, generated from the 1000 eigenvalues of the sample covariance matrix for 1000 standard normal random vectors in $\mathbb{R}^{4000}$. In this data set, $\lambda_1 = 8.9517$ and $\lambda_{1000} = 1.0145$; we will see later in this chapter that, for large N, with N data points in $\mathbb{R}^m$ where $m/N \sim 4$, the histogram is supported on $[1, 9]$.

A histogram $h_{\Pi,\boldsymbol{\lambda}}$ is a probability density function (under our definition which normalizes with a $\frac{1}{N}$). Indeed: it is a step function which takes possible values $\frac{k}{N}$ where $k = 0, 1, 2, \ldots, N$, and so it is always ≥ 0. The integral $\int_{\mathbb{R}} h_{\Pi,\boldsymbol{\lambda}}(t)\, dt$ is the sum of the values on all of the partition intervals; each bin $(t_i - 1, t_i]$ contains some number k_i of the points from $\boldsymbol{\lambda}$, and so the sum of the values is $\sum_i \frac{k_i}{N} = \frac{1}{N}\sum_i k_i = \frac{1}{N}(N) = 1$, since the k_i add up to the total number of points, N. (This is why it's important not to double count which bin any point is in.) It can be thought of as an *approximate density* for the data $\boldsymbol{\lambda}$. To be precise:

DEFINITION 2.4. *Let $\boldsymbol{\lambda}$ be a point set of finite size N. The associated* **empirical random variable** *$E_{\boldsymbol{\lambda}}$ is a discrete random variable, whose distribution is defined by*

$$\mathbb{P}(E_{\boldsymbol{\lambda}} = x) = \begin{cases} \frac{1}{N} & \text{if } x = \lambda_j \text{ for some } j \in \{1, \ldots, N\} \\ 0 & \text{if } x \notin \boldsymbol{\lambda}. \end{cases}$$

To be persnickety: it could be that two or more of the λ_j are equal; in that case, a more careful definition would assign probability $\frac{k}{N}$ to that point, where k is the number of labels i for which λ_i coincides. A fully consistent definition of the distribution of $E_{\boldsymbol{\lambda}}$ is

$$\mathbb{P}(E_{\boldsymbol{\lambda}} \in A) = \frac{1}{N}\#(A \cap \boldsymbol{\lambda}) = \frac{1}{N}\#\{j \colon \lambda_j \in A\}. \tag{2.2}$$

The empirical random variable $E_{\boldsymbol{\lambda}}$ is discrete; it definitely does not have a probability density. Any histogram $h_{\Pi,\boldsymbol{\lambda}}$ is a way to "smear out" the probability masses, to describe its distribution (approximately) via a probability density function. Following this line of reasoning (i.e. wishful thinking), suppose that $E_{\boldsymbol{\lambda}}$ really *did* have a probability density function $f_{\boldsymbol{\lambda}}$. Then we could compute probabilities for $E_{\boldsymbol{\lambda}}$ using this density in the usual way: for any fixed interval $[a, b]$,

$$\mathbb{P}(E_{\boldsymbol{\lambda}} \in [a, b]) = \int_a^b f_{\boldsymbol{\lambda}}(\lambda)\, d\lambda$$

and combining this with (2.2), this would mean

$$\frac{1}{N}\#\{i\colon \lambda_i \in [a,b]\} = \mathbb{P}(E_{\boldsymbol{\lambda}} \in [a,b]) = \int_a^b f_{\boldsymbol{\lambda}}(\lambda)\, d\lambda. \tag{2.3}$$

There is, of course, no such density $f_{\boldsymbol{\lambda}}$; nevertheless, it is useful to think of this fantasy object as we motivate our discussion in the next sections, where we will see that such a density *does* emerge as the point set $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(N)}$ grows with N (in our random matrix theory setting).

The well-defined quantities on the left-hand-side of (2.3) actually encode all the information we need to construct any histogram of the points. Indeed, the function $h_{\Pi,\lambda}$ is piecewise constant; if $(t_{i-1}, t_i]$ is one of the partition bins in Π, then $h_{\Pi,\lambda}(t) = \frac{1}{N}\#\{j\colon \lambda_j \in (t_{i-1}, t_i]\}$. Hence, all we really need to keep track of are the quantities:

$$h_{[a,b]}(\boldsymbol{\lambda}) = \frac{1}{N}\#\{j\colon \lambda_j \in [a,b]\}, \qquad \text{for any } \ a < b. \tag{2.4}$$

REMARK 2.5. To be persnickety, we should be consistent about bins being of the form $(t_{i-1}, t_i]$ and not $[t_{i-1}, t_i]$ to avoid overcounting at the partition points (i.e. if one of the λ_j happens to be at a break point t_i for the bins). Nevertheless, it is still enough just to know the numbers $h_{[a,b]}(\boldsymbol{\lambda})$ in (2.4), since the interval $(t_{i-1}, t_i]$ is the union of all closed intervals of the form $[t_{i-1} + \epsilon, t_i]$ for $\epsilon \downarrow 0$; hence, we can recover the proportion of $\boldsymbol{\lambda}$ points in $(t_{i-1}, t_i]$ as $\lim_{\epsilon\downarrow 0} h_{[t_{i-1}+\epsilon, t_i]}(\boldsymbol{\lambda})$ — a limit of numbers all of the form (2.4).

We are therefore interested in computing the counting random variables $h_{[a,b]}(\boldsymbol{\lambda})$ when $\boldsymbol{\lambda}$ are the eigenvalues of a random matrix. As with any counting random variable, a useful trick will be the *method of indicators*: we want to express this random number as a sum of Bernoulli (i.e. $\{0,1\}$-valued) random variables that we can analyze more easily. In this case, there is an immediate solution:

$$h_{[a,b]}(\boldsymbol{\lambda}) = \frac{1}{N}\#\{j\colon \lambda_j \in [a,b]\} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{[a,b]}(\lambda_j). \tag{2.5}$$

Indeed: $\mathbb{1}_{[a,b]}$, the indicator function of the interval $[a,b]$, takes values $\mathbb{1}_{[a,b]}(\lambda) = 1$ if $\lambda \in [a,b]$ and $\mathbb{1}_{[a,b]}(\lambda) = 0$ if $\lambda \notin [a,b]$. Thus, the sum in (2.5) is composed of terms that are all 0 or 1, and so adds up to the number of terms where the value is 1; i.e. the number of terms where $\lambda_j \in [a,b]$, as desired.

What have we gained by doing this? Without direct knowledge of the distributions of the λ_j, how is this sum going to help us? The answer lies in expanding our horizons and considering a larger family of such summations, some of which will be computable (as we'll see in the next section).

DEFINITION 2.6. *Let $\boldsymbol{\lambda}$ be a set of N points $\lambda_1 \geq \cdots \geq \lambda_N$ in $\mathbb{R}$. A* **linear statistic** *of $\boldsymbol{\lambda}$ is a quantity of the form*

$$L(\boldsymbol{\lambda}) = \frac{1}{N}\sum_{j=1}^{N} w(\lambda_j)$$

for some "weight" function $w\colon \mathbb{R} \to \mathbb{R}$.

The histogram statistics (2.4) are thus linear statistics, by (2.5), with the function being $w = \mathbb{1}_{[a,b]}$. We have already been working with some linear statistics extensively, to be clear. The **sample mean** $\overline{\lambda}_N = \frac{1}{N}\sum_{j=1}^{N} \lambda_j$ is a linear statistic, using $w(\lambda) = \lambda$. The sample variance

$$S_N(\boldsymbol{\lambda}) = \frac{1}{N}\sum_{j=1}^{N}(\lambda_j - \overline{\lambda}_N)^2 = \frac{1}{N}\sum_{j=1}^{N}\lambda_j^2 - (\overline{\lambda}_N)^2$$

is not *quite* a linear statistic, since it involves the square of the sample mean, and the square of a linear statistic is not generally a linear statistic. However, the non-centered version $\frac{1}{N}\sum_{j=1}^{N}\lambda_j^2$ *is* a linear statistic, with $w(\lambda) = \lambda^2$. In general, the **sample moments**

$$M_k(\boldsymbol{\lambda}) = \frac{1}{N}\sum_{j=1}^{N}(\lambda_j)^k, \qquad k \in \mathbb{N} \tag{2.6}$$

are all linear statistics (using $w(\lambda) = \lambda^k$). These particular linear statistics will play an important role in our analysis.

This is well and good, but how will expanding out horizons to include all kinds of linear statistics help us with the original goal: computing the histogram linear statistics $h_{[a,b]}(\boldsymbol{\lambda})$? The answer is, amazingly, that the very restricted class of sample moments (2.6) are, by themselves, *enough to determine all linear statistics* of a finite point set. The reason is the following approximation lemma, which is due to Russian mathematician Sergei Natanovich Bernstein.

LEMMA 2.7 (Bernstein, 1920s). *Let $[\alpha, \beta]$ be a closed and bounded interval, and let $w\colon [\alpha,\beta] \to \mathbb{R}$ be a function. There is a sequence of polynomials $\{B_n\}_{n\in\mathbb{N}}$, where B_n has degree n, such that $\lim_{n\to\infty} B_n(t) = w(t)$ for all points $t \in [\alpha,\beta]$ where w is continuous.*

PROOF. We give a version of Bernstein's original proof, which is ingenious in its use of what is now called "the probabilistic method": introduce random variables where there weren't any obvious, to leverage the power of the theorems of probability theory.

First, we note that it suffices to prove the theorem in the case $\alpha = 0$ and $\beta = 1$. This is because we can, in general, do an affine transformation of the the argument of the function $t \mapsto (\beta-\alpha)t+\alpha$ to transform it into a new function defined on $[0,1]$. If we can prove the desired convergence property for this transformed function and find the approximating polynomials B_n, we can then transform both the function and the polynomials back via the inverse transform $t \mapsto \frac{t-\alpha}{\beta-\alpha}$; this transform preserves polynomials (of degree n) and continuity, and it is straightforward to check that the statement of the lemma is maintained under this transformation.

Hence, we proceed to prove the lemma in the case that the domain of the function is the interval $[0,1]$; we will denote the argument of the function as p, as we are going to interpret it as a probability! Indeed, let $S_{n,p}$ denote a Binomial random with parameters n and p; that is, let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables with expectation p, and set $S_{n,p} = X_1 + \cdots + X_n$. By the strong law of large numbers,

$$\frac{1}{n}S_{n,p} = \frac{1}{n}\sum_{j=1}^{n}X_j \to p \text{ with probability } 1 \text{ as } n \to \infty.$$

It therefore follows that, if w is continuous at p, then

$$\lim_{n\to\infty} w\big(\tfrac{1}{n}S_{n,p}\big) = w(p) \text{ with probability } 1.$$

Because this is almost sure convergence, it certainly implies convergence of the expectation; thus

$$\lim_{n\to\infty} \mathbb{E}\left[w\big(\tfrac{1}{n}S_{n,p}\big)\right] = w(p). \tag{2.7}$$

We can now readily compute this expectation. The random variable $\frac{1}{n}S_{n,p}$ is discrete, taking possible values $0, \frac{1}{n}, \frac{2}{n}, \dots, 1$, where $\mathbb{P}(\frac{1}{n}S_{n,p} = \frac{k}{n}) = \binom{n}{k}p^k(1-p)^{n-k}$. Hence

$$\mathbb{E}\left[w\left(\tfrac{1}{n}S_{n,p}\right)\right] = \sum_{k=0}^{n} w\left(\tfrac{k}{n}\right)\binom{n}{k}p^k(1-p)^{n-k} =: B_n(p).$$

This is a polynomial in the variable p, of degree n. We call this polynomial $B_n(p)$ the *Bernstein polynomial* of w (on $[0,1]$). Equation (2.7) shows that $B_n(p) \to w(p)$ at all points p where f is continuous, concluding the proof. □

REMARK 2.8 (This remark can be safely ignored; it includes some fun tid-bits for those who have taken MATH 140B). A slightly more involved version of this same argument shows that, if w is continuous on the whole compact interval, the convergence $B_n \to w$ is actually uniform. This gave the first known *constructive* proof of the Weierstraß approximation theorem. More generally, the very same approximation procedure shows that the Bernstein polynomials of any *measurable* function converge to that function (on a compact interval) almost everywhere.

We can use Bernstein polynomials to approximate the indicator function $w = \mathbb{1}_{[a,b]}$, on some larger interval. For example: suppose we know that our random points $\boldsymbol{\lambda}$ are all in the interval $[\alpha, \beta]$. Then for $\alpha < a < b < \beta$, we can construct the Bernstein polynomials B_n of the function $\mathbb{1}_{[a,b]}$ in the interval $[\alpha, \beta]$; by Lemma 2.7, $B_n(t) \to \mathbb{1}_{[a,b]}(t)$ for all $t \in [\alpha, \beta]$ except possibly $t = a$ and $t = b$. That is: $B_n(t) \to 1$ if $t \in (a, b)$ and $B_n(t) \to 0$ if $t \in [\alpha, a) \cup (b, \beta]$. Therefore, so long as none of the points in $\boldsymbol{\lambda}$ are at a or b, we have

$$B_n(\lambda_j) \to \mathbb{1}_{[a,b]}(\lambda_j) \quad \text{as} \quad n \to \infty$$

and hence

$$\frac{1}{N}\sum_{j=1}^{N} B_n(\lambda_j) \to \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{[a,b]}(\lambda_j) = h_{[a,b]}(\boldsymbol{\lambda}) \quad \text{as} \quad n \to \infty. \tag{2.8}$$

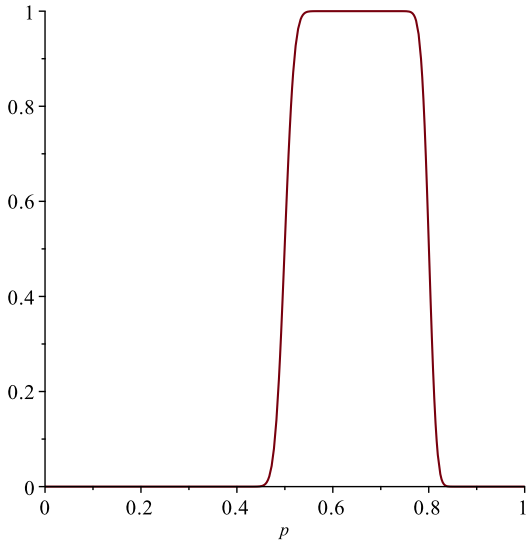


FIGURE 2.2. The Bernstein polynomial B_{1000} of the indicator function $\mathbb{1}_{[0.5,0.8]}$ in the unit interval $[0,1]$.

REMARK 2.9. (1) The polynomials B_n cannot possibly continue to approximate $\mathbb{1}_{[a,b]}$ on arbitrarily large intervals, because all non-constant polynomials must diverge to $\pm\infty$ at the "ends" of $\mathbb{R}$. But that is not a problem for us; since we have chosen the large interval $[\alpha, \beta]$ to contain all the points $\boldsymbol{\lambda}$ that we care about, it doesn't matter what happens outside this interval.

(2) The above discussion presents problems in the special case that one or more of the λ_j happens to fall at an endpoint a, b of the interval in question. This can be gotten around by "jiggling" the interval a little bit, cf. Remark 2.5. Alternatively: in the cases of interest to us, $\boldsymbol{\lambda}$ will be a *random* point set, which will possess a joint density; this means that the probability of any of the λ_j taking any one of the finitely many partition points will always be 0, so we can safely ignore this problem which will almost never come up.

The conclusion of all this is that *it suffices to compute polynomial linear statistics in order to compute histogram statistics of any finite points set*. In fact, we can go one step further.

PROPOSITION 2.10. *Let $\boldsymbol{\lambda}$ be a finite set of N points in $\mathbb{R}$. Suppose we can calculate all of the sample moments $M_k(\boldsymbol{\lambda})$ for $k \in \mathbb{N}$ (cf.* (2.6)*). Then we can compute the histogram statistics $h_{[a,b]}(\boldsymbol{\lambda})$ for all intervals $[a, b]$.*

PROOF. Fix a large interval $[\alpha, \beta]$ that contains all the points in $\boldsymbol{\lambda}$. Compute the Bernstein polynomials B_n of the function $\mathbb{1}_{[a,b]}$ in this larger interval, cf. Lemma 2.7. The Bernstein polynomial B_n is a polynomial of degree n, and so it can be written out as

$$B_n(\lambda) = \sum_{k=0}^{n} b_{n,k}\lambda^k$$

for some coefficients $b_{n,k}$. The corresponding linear statistics are

$$\frac{1}{N}\sum_{j=1}^{N} B_n(\lambda_j) = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=0}^{n} b_{n,k}(\lambda_j)^k = \sum_{k=0}^{n} b_{n,k}\frac{1}{N}\sum_{j=1}^{N}(\lambda_j)^k = \sum_{k=0}^{n} b_{n,k}M_k(\boldsymbol{\lambda}). \tag{2.9}$$

Now, the B_n were constructed so the left-hand-side of (2.9) converges to the histogram statistic $h_{[a,b]}(\boldsymbol{\lambda})$ as $n \to \infty$, cf. (2.8). Hence, we can compute the histogram statistics as limits of terms on the right-hand-side of (2.9), which are linear combinations of sample moments. Therefore, if we know how to compute sample moments, we know how to compute histogram statistics. □

REMARK 2.11. The sense of "knowing how to compute" in the above proposition is very theoretical; although it could be made quantitatively precise (in terms of how large n must be taken to compute $h_{[a,b]}(\boldsymbol{\lambda})$ to desired accuracy), we would never actually use it to compute a histogram (which, numerically, is a triviality to produce). Rather, the use of this proposition is theoretical: it tells us that if we can compute *and recognize* the sample moments of the points as the actual moments of some probability density function, then we can conclude that density underlies the histogram of the points — it is the "true" density alluded to in the discussion following Figure 2.1. We will make this more precise in the next section.

2.2. Convergence of Sample Moments

We are interested in the properties of histograms of the eigenvalues $\boldsymbol{\lambda}$ of an $N \times N$ random matrix. Since the entries of the matrix are random, so are its eigenvalues. From the examples in Figure 2.1, we can see that the histograms are also going to be random; nevertheless, there is some apparent underlying deterministic structure, and that is what we'd like to understand. Motivated

by the limit theorems of probability theory, our hunch is that the random fluctuations in those histograms will get smaller as N grows, and so what we are looking for is a description of the **large-N limit histograms of eigenvalues**. For this reason, as we will have to allow N to grow, we will label the eigenvalue point set with N:

$$\boldsymbol{\lambda}^{(N)} = \{\lambda_1^{(N)}, \lambda_2^{(N)}, \dots, \lambda_N^{(N)}\}$$

still with the convention that $\lambda_1^{(N)} \geq \lambda_2^{(N)} \geq \cdots \geq \lambda_N^{(N)}$.

Our goal is to understand, for each fixed finite interval $[a, b]$, the large-N limit of the histogram statistic $h_{[a,b]}(\boldsymbol{\lambda}^{(N)})$:

$$\lim_{N\to\infty} h_{[a,b]}(\boldsymbol{\lambda}^{(N)}) = \lim_{N\to\infty} \frac{1}{N}\#\{j\colon \lambda_j^{(N)} \in [a,b]\}. \tag{2.10}$$

Since we have little-to-no information about the eigenvalues $\boldsymbol{\lambda}^{(N)}$ directly, we cannot say much directly about these quantities. However, appealing to Proposition 2.10, we can (in principle) get information about histogram statistics just from computing the sample moments:

$$M_k(\boldsymbol{\lambda}^{(N)}), \quad k \in \mathbb{N}.$$

Proposition 2.10 was stated in the context of a fixed point set $\boldsymbol{\lambda}$; now we will be dealing with a changing point set $\boldsymbol{\lambda}^{(N)}$ (with a growing number of points). Since the recovery of $h_{[a,b]}$ from M_k involved a limiting approximation procedure, we need to be very careful now, since there will be two simultaneous limits. This section is devoted to clarifying these points.

Keeping our eye on the prize: what sort of structure to we hope to see? What do we expect to emerge as a description of the large-N limit in (2.10)? Recall that any histogram (normalized in our convention) is a probability density, which is supposed to represent an "approximate density" for the discrete *empirical random random variable* $E_{\boldsymbol{\lambda}^{(N)}}$, cf. Definition 2.4. Our idealized model would see a true density $f^{(N)}$ for the random variable $E_{\boldsymbol{\lambda}^{(N)}}$, as in (2.3):

$$h_{[a,b]}(\boldsymbol{\lambda}^{(N)}) = \int_a^b f^{(N)}(\lambda)\, d\lambda.$$

This is not possible, but we will see what *does* happen is that such a density emerges in the large-N limit.

DEFINITION 2.12. *Let $\boldsymbol{\lambda}^{(N)} \subset \mathbb{R}$ be point sets of size N; we call such a sequence of point sets an* **ensemble**. *An ensemble has an* **asymptotic probability density** *f if, for every $a < b$,*

$$\lim_{N\to\infty} h_{[a,b]}(\boldsymbol{\lambda}^{(N)}) = \int_a^b f(\lambda)\, d\lambda.$$

We are going to prove that, when the ensemble $\boldsymbol{\lambda}^{(N)}$ consists of the (random) eigenvalues of the sample covariance matrix of truly random data, it does indeed have an asymptotic probability density, which we see emerging in the histograms of Figure 2.1; and we will compute this density. To do so, we need to have some access to the (large-N limits of the) histogram random variables $h_{[a,b]}(\boldsymbol{\lambda}^{(N)})$ using tools we can compute with. To that end, Proposition 2.10 is the key. Here is the main theorem.

THEOREM 2.13. *Let $(\boldsymbol{\lambda}^{(N)})_{N\in\mathbb{N}}$ be an ensemble of finite point sets in $\mathbb{R}$, and let $(M_k(\boldsymbol{\lambda}^{(N)}))_{k\in\mathbb{N}}$ be the sample moments (cf.* (2.6)*). Suppose that, for each k,*

$$\lim_{N\to\infty} M_k(\boldsymbol{\lambda}^{(N)}) = \mu_k$$

exists, and that there is a probability density f, supported on a bounded interval $[\alpha, \beta]$, whose moments are $(\mu_k)_{k\in\mathbb{N}}$:

$$\int_{\mathbb{R}} \lambda^k f(\lambda)\, d\lambda = \mu_k.$$

Then the ensemble has f as an asymptotic probability density.

PROOF. We are going to provide the outline of the proof, but will assert a key technical fact whose proof (while within our reach) would take *much* more time and energy, and would really obscure the basic idea.

Fix $a < b$, and let (B_n) be the Bernstein polynomials approximating $\mathbb{1}_{[a,b]}$ on the interval $[\alpha, \beta]$, cf. Proposition 2.10. Denote its coefficients as $b_{n,k}$, so $B_n(\lambda) = \sum_{k=0}^n b_{n,k}\lambda^k$. Note that

$$\sum_{k=0}^n b_{n,k} M_k(\boldsymbol{\lambda}^{(N)}) = \sum_{k=0}^n b_{n,k} \frac{1}{N}\sum_{j=1}^N (\lambda_j^{(N)})^k = \frac{1}{N}\sum_{j=1}^N \sum_{k=0}^n b_{n,k}(\lambda_j^{(N)})^k = \frac{1}{N}\sum_{j=1}^N B_n(\lambda_j^{(N)}).$$

Now, if we send $n \to \infty$, since $B_n(\lambda) \to \mathbb{1}_{[a,b]}(\lambda)$ except possibly if $\lambda \in \{a, b\}$, we conclude that (for almost all choices of a, b)

$$\lim_{n\to\infty} \sum_{k=0}^n b_{n,k} M_k(\boldsymbol{\lambda}^{(N)}) = \frac{1}{N}\sum_{j=1}^N \lim_{n\to\infty} B_n(\lambda_j^{(N)}) = \frac{1}{N}\sum_{j=1}^N h_{[a,b]}(\lambda_j^{(N)}) = h_{[a,b]}(\boldsymbol{\lambda}^{(N)}).$$

Our goal is to show that $\lim_{N\to\infty} h_{[a,b]}(\boldsymbol{\lambda}^{(N)}) = \int_a^b f(\lambda)\, d\lambda$ (this is the definition of "asymptotic probability density" for an ensemble, cf. Definition 2.12), so we should try to compute this limit:

$$\lim_{N\to\infty} h_{[a,b]}(\boldsymbol{\lambda}^{(N)}) = \lim_{N\to\infty}\lim_{n\to\infty} \sum_{k=0}^n b_{n,k} M_k(\boldsymbol{\lambda}^{(N)}). \tag{2.11}$$

Now we make the technical assertion that we will not prove:

$$\lim_{N\to\infty}\lim_{n\to\infty} \sum_{k=0}^n b_{n,k} M_k(\boldsymbol{\lambda}^{(N)}) = \lim_{n\to\infty}\lim_{N\to\infty} \sum_{k=0}^n b_{n,k} M_k(\boldsymbol{\lambda}^{(N)}). \tag{2.12}$$

(This is no simple matter in general: reversing the order of two limits can be very tricky to justify, and is typically just not true. It turns out to be true in our case; this is where the assumption of the bounded support of the limit density f comes into play.) With this in hand, we compute the N-limit holding n fixed:

$$\lim_{N\to\infty} \sum_{k=0}^n b_{n,k} M_k(\boldsymbol{\lambda}^{(N)}) = \sum_{k=0}^n b_{n,k} \lim_{N\to\infty} M_k(\boldsymbol{\lambda}^{(N)}) = \sum_{k=0}^n b_{n,k}\mu_k$$

utilizing the assumptions of the theorem. Moreover, the assumptions further imply that these μ_k are the moments of f, so the above quantity equals

$$\sum_{k=0}^n b_{n,k} \int_{\mathbb{R}} \lambda^k f(\lambda)\, d\lambda = \int_{\mathbb{R}} \sum_{k=0}^n b_{n,k}\lambda^k f(\lambda)\, d\lambda = \int_{\mathbb{R}} B_n(\lambda) f(\lambda)\, d\lambda.$$

Thus, from (2.11) and (2.12), we have

$$\lim_{N\to\infty} h_{[a,b]}(\boldsymbol{\lambda}^{(N)}) = \lim_{n\to\infty} \int_{\mathbb{R}} B_n(\lambda) f(\lambda)\, d\lambda = \lim_{n\to\infty} \int_\alpha^\beta B_n(\lambda) f(\lambda)\, d\lambda$$

where the last equality follows from the fact (in the assumption of the theorem) that f is supported on $[\alpha, \beta]$. Finally, we must move the limit inside the integral. This, again, takes some work; in this case, it is not so bad because B_n approximates the bounded function $\mathbb{1}_{[a,b]}$ on $[\alpha, \beta]$, and so this limit interchange can be justified using the bounded convergence theorem. The result is

$$\lim_{N\to\infty} h_{[a,b]}(\boldsymbol{\lambda}^{(N)}) = \int_\alpha^\beta \lim_{n\to\infty} B_n(\lambda) f(\lambda)\, d\lambda = \int_\alpha^\beta \mathbb{1}_{[a,b]}(\lambda) f(\lambda)\, d\lambda = \int_a^b f(\lambda)\, d\lambda.$$

(By construction, i.e. Proposition 2.10, $\lim_{n\to\infty} B_n(\lambda) = h_{[a,b]}(\lambda)$ at all points other than possibly $\lambda = a, b$; since the integral doesn't care what happens at any finite number of points, this convergence is good enough for what we asserted here.) □

REMARK 2.14. More on the magic behind the limit interchange (2.12): one useful criterion that allows limits to be interchanged is *uniformity*. If we knew, for example, that the n-limit was uniform in N (meaning the rate of convergence is independent of N), then an off-the-shelf theorem would justify the limit interchange. Unfortunately, this is not true in the present context, because B_n cannot converge uniformly to the discontinuous function $\mathbb{1}_{[a,b]}$. For that reason, this theorem is often proved in two steps: first, we follow the above outline to show how to recover any asymptotic linear statistic with a *continuous* weight function w, since these can be uniformly approximated by Bernstein polynomials; then approximate $\mathbb{1}_{[a,b]}$ by continuous functions after the limit interchange. We could carry out this procedure without any more tools than we have now, but it would take several pages of work, and would really obscure the core idea of the proof.

The beauty of Theorem 2.13 is that we can now safely identify the asymptotic probability density so long as we can compute sample moments. More precisely, our process will be:

(1) Compute the sample moments M_k of our eigenvalue ensemble.
(2) Show that they have large-N limits μ_k, and compute these limit moments.
(3) Identify the probability density that has these moments.
(4) Profit.

The next section discusses how we will handle items (1) and (2).

2.3. Convergence and Concentration of Moments

We now turn to the specific ensembles of interest: the eigenvalues of certain random matrices. Still staying somewhat general, let's consider a general $N \times N$ symmetric random matrix W_N; its eigenvalues will be denoted $\boldsymbol{\lambda}^{(N)} = \{\lambda_1^{(N)}, \dots, \lambda_N^{(N)}\}$ (where, as usual, our convention is that $\lambda_1^{(N)} \geq \lambda_2^{(N)} \geq \cdots \geq \lambda_N^{(N)}$). We have seen above that, to understand the large-N limit histogram statistics of this ensemble of eigenvalues, it suffices to compute the sample moments $M_k(\boldsymbol{\lambda}^{(N)})$, and, in particular, their large-N limits. We know that computing eigenvalues (exactly, analytically) is a hopeless task; fortunately, we do not need to compute the eigenvalues in order to compute all their sample moments.

PROPOSITION 2.15. *Let W_N be an $N \times N$ symmetric matrix with eigenvalues $\boldsymbol{\lambda}^{(N)}$. The sample moments of the eigenvalues can be computed as*

$$M_k(\boldsymbol{\lambda}^{(N)}) = \frac{1}{N} \operatorname{Tr}\left[(W_N)^k\right].$$

To prove this, we need the following lemma, which is the key to why the trace is a useful linear functional on matrices.

LEMMA 2.16. *Let $N, m \in \mathbb{N}$, and let $A \in \mathbb{M}_{Nm}$ and $B \in \mathbb{M}_{mN}$ be two matrices. Then*

$$\operatorname{Tr}(AB) = \operatorname{Tr}(BA).$$

PROOF. By definition, for any square matrix C, $\operatorname{Tr}(C) = \sum_i [C]_{ii}$ is the sum of the diagonal entries. We now compute

$$[AB]_{ii} = \sum_{j=1}^{m} [A]_{ij}[B]_{ji}$$

and so

$$\operatorname{Tr}(AB) = \sum_{i=1}^{N} [AB]_{ii} = \sum_{i=1}^{N}\sum_{j=1}^{m} [A]_{ij}[B]_{ji} = \sum_{j=1}^{m}\sum_{i=1}^{N} [A]_{ij}[B]_{ji}.$$

Having reversed the order of the summations, we note that the internal sum is

$$\sum_{i=1}^{N} [A]_{ij}[B]_{ji} = \sum_{i=1}^{N} [B]_{ji}[A]_{ij} = [BA]_{jj}$$

and so the double sum is also equal to $\sum_j [BA]_{jj} = \operatorname{Tr}(BA)$. □

REMARK 2.17. We can apply this lemma to longer products, one step at a time; for example

$$\operatorname{Tr}(ABC) = \operatorname{Tr}((AB)C) = \operatorname{Tr}(C(AB)) = \operatorname{Tr}(CAB).$$

Following this reasoning, we see that the trace is invariant under any *cyclic* permutation of a product of matrices. **Caution**: it is *not* invariant under other permutations. For example, while $\operatorname{Tr}(ABC) = \operatorname{Tr}(CAB) = \operatorname{Tr}(BCA)$, in general $\operatorname{Tr}(ABC) \neq \operatorname{Tr}(BAC)$.

PROOF OF PROPOSITION 2.15. Since W_N is symmetric, by the spectral theorem, it can be (orthogonally) diagonalized $W_N = Q_N \Lambda_N Q_N^{-1}$. Hence, taking powers,

$$\begin{aligned}(W_N)^k &= (Q_N\Lambda_N Q_N^{-1})(Q_N\Lambda_N Q_N^{-1})\cdots(Q_N\Lambda_N Q_N^{-1}) \\ &= Q_N\Lambda_N(Q_N Q_N^{-1})\Lambda_N(Q_N Q_N^{-1})\Lambda_N\cdots\Lambda_N(Q_N Q_N^{-1})\Lambda_N Q_N^{-1} \\ &= Q_N(\Lambda_N)^k Q_N^{-1}.\end{aligned}$$

Now, utilizing Lemma 2.16, it follows that

$$\operatorname{Tr}[(W_N)^k] = \operatorname{Tr}[Q_N(\Lambda_N)^k Q_N^{-1}] = \operatorname{Tr}[Q_N^{-1}Q_N(\Lambda_N)^k = \operatorname{Tr}[(\Lambda_N)^k].$$

Finally: Λ_N is a diagonal matrix, with the eigenvalues on the diagonal; its powers are thus

$$(\Lambda_N)^k = \begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N\end{bmatrix}^k = \begin{bmatrix}(\lambda_1)^k & & & \\ & (\lambda_2)^k & & \\ & & \ddots & \\ & & & (\lambda_N)^k\end{bmatrix}$$

and the trace of this diagonal matrix is just $\sum_j (\lambda_j)^k$. (Here we have suppressed the extra index $\lambda_j = \lambda_j^{(N)}$ for readability.) Thus

$$\operatorname{Tr}[(W_N)^k] = \sum_{j=1}^{N} (\lambda_j)^k = N \cdot M_k(\boldsymbol{\lambda}^{(N)})$$

as desired. □

The amazing point here is that the quantities $M_k(\boldsymbol{\lambda}^{(N)})$ which are explicit functions of the eigenvalues $\lambda_j^{(N)}$ can all be calculated directly and only from the *entries* of the matrix W_N: the entries of $(W_N)^k$ are homogeneous degree k polynomials in the entires of W_N. Let's look at an example.

EXAMPLE 2.18. Let W_N be the kind of sample covariance matrix we are interested in studying: $W_N = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$ where $\mathbf{X} \in \mathcal{M}_{m\times N}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Then

$$[W_N]_{ii} = \frac{1}{N}\sum_{j=1}^{m}[\mathbf{X}^\top]_{ij}[\mathbf{X}]_{ji} = \frac{1}{N}\sum_{j=1}^{m}[\mathbf{X}]_{ji}^2$$

and therefore the sample mean of the eigenvalues of W is equal to

$$M_1(\boldsymbol{\lambda}^{(N)}) = \frac{1}{N}\operatorname{Tr}(W_N) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{m}[\mathbf{X}]_{ji}^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{m}[\mathbf{X}]_{ji}^2. \tag{2.13}$$

The Nm random variables $[\mathbf{X}]_{ji}$ are all i.i.d. standard normals, so $M_1(\boldsymbol{\lambda}^{(N)})$ is $\frac{1}{N^2}$ times a χ^2 random variable with Nm degrees of freedom.

However, since Nm is large, this random variable is quite concentrated about its mean. Indeed: the $[\mathbf{X}]_{ji}^2$ are i.i.d., with common expectation $\mathbb{E}([\mathbf{X}]_{ji}^2) = 1$, and so by the Laws of Large Numbers,

$$\frac{1}{Nm}\sum_{i=1}^{N}\sum_{j=1}^{m}[\mathbf{X}]_{ji}^2 \to 1 \qquad \text{as } \ Nm \to \infty.$$

We computed above that

$$M_1(\boldsymbol{\lambda}^{(N)}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{m}[\mathbf{X}]_{ji}^2 = \frac{m}{N}\cdot\frac{1}{Nm}\sum_{i=1}^{N}\sum_{j=1}^{m}[\mathbf{X}]_{ji}^2$$

and so we see the following: *if $N, m \to \infty$ in such a way that the ratio $\frac{m}{N} \to \varrho \in (0,\infty)$*, then

$$\mu_1 = \lim_{N\to\infty} M_1(\boldsymbol{\lambda}^{(N)}) = \varrho.$$

REMARK 2.19. The calculation of (2.13) shows that for any rectangular matrix $\mathbf{X}$, $\operatorname{Tr}(\mathbf{X}^\top\mathbf{X}) = \sum_{i,j}[\mathbf{X}]_{ji}^2$ is the sum of the squares of all the entries. In other words: if we think if the rectangular matrix $\mathbf{X}$ as being a Euclidean vector (think about reading down the columns of the matrix one by one into one long column), then its length^2 can be computed as $\operatorname{Tr}(\mathbf{X}^\top\mathbf{X})$. This convenient connection between Euclidean geometry and the algebraic properties of matrices (involving transpose, product, and trace) goes a long way to explaining why there is such nice structure in the statistics of the eigenvalues. The quantity $(\operatorname{Tr}(\mathbf{X}^\top\mathbf{X}))^{1/2}$ is called the **Hilbert–Schmidt norm** (or sometimes the **Fröbenius norm**) of the matrix.

We will proceed in the next section to compute higher sample moments of the random matrix W_N in Example 2.18. In the case of the sample mean M_1, the formula is simple enough to completely describe the distribution: a rescaled χ^2. For higher sample moments, the formulas will be increasingly complicated and it will be very difficult to compute and describe the distribution exactly. Nevertheless, since we are primarily interested in the large-N limit, we will see that a phenomenon like in Example 2.18 will always come into play: we will have *concentration around the mean*. The way this will manifest is as follows.

PROPOSITION 2.20. *Let $(\boldsymbol{\lambda}^{(N)})_{N\in\mathbb{N}}$ be an ensemble of random points, and let $M^{(N)}$ be any statistic of $(\boldsymbol{\lambda}^{(N)})$ (e.g. $M^{(N)} = M_k(\boldsymbol{\lambda}^{(N)})$ for some $k \in \mathbb{N}$). Suppose that the following two conditions holds for the (random) sample moments of the ensemble:*

(1) $\lim_{N\to\infty} \mathbb{E}[M^{(N)}] = \mu$ *exists, and*
(2) $\lim_{N\to\infty} \mathrm{Var}[M^{(N)}] = 0$.

Then $M^{(N)}$ converges in probability to μ as $N \to \infty$.

PROOF. This is basically the same argument as the proof of the Weak Law of Large Numbers (using Chebyshev's inequality); it just requires a slight tweak because μ is not necessarily *equal* to the expected value of $M^{(N)}$, but it is the *limit* of these expected values. First, we have

$$|M^{(N)} - \mu]| \leq |M^{(N)} - \mathbb{E}[M^{(N)}]| + |\mathbb{E}[M^{(N)}] - \mu| \tag{2.14}$$

by the triangle inequality. Hence, for any $\epsilon > 0$,

$$\{|M^{(N)} - \mu| > \epsilon\} \subseteq \{|M^{(N)} - \mathbb{E}[M^{(N)}]| > \epsilon/2\} \cup \{|\mathbb{E}[M^{(N)}] - \mu| > \epsilon/2\}.$$

(Indeed: if $|M^{(N)} - \mu| > \epsilon$ then, by (2.14), $|M^{(N)} - \mathbb{E}[M^{(N)}]| + |\mathbb{E}[M^{(N)}] - \mu| > \epsilon$, and if this sum of two positive terms is $> \epsilon$ then at least one of them must be $> \epsilon/2$.) Therefore

$$\mathbb{P}(|M^{(N)} - \mu| > \epsilon) \leq \mathbb{P}(|M^{(N)} - \mathbb{E}[M^{(N)}]| > \epsilon/2) + \mathbb{P}(|\mathbb{E}[M^{(N)}] - \mu| > \epsilon/2). \tag{2.15}$$

By Chebyshev's inequality, the first term in (2.15) is bounded as follows:

$$\mathbb{P}(|M^{(N)} - \mathbb{E}[M^{(N)}]| > \epsilon/2) \leq \frac{\mathrm{Var}[M^{(N)}]}{(\epsilon/2)^2}.$$

By assumption (2), the right-hand-side tends to 0 as $N \to \infty$. As for the second term in (2.15), the event in question is not random: for each N, $|\mathbb{E}[M^{(N)}] - \mu|$ is a constant. By assumption (1), this constant tends to 0, which means that for the given ϵ, this constant is $\leq \epsilon/2$ for all large N. Thus, the second term in (2.15) is actually identically 0 for all large N. Combining these observations, wee conclude that $M^{(N)} \to_{\mathbb{P}} \mu$ (i.e. the distribution of $M^{(N)}$ concentrates around μ as $N \to \infty$) as claimed. □

REMARK 2.21. There is a seminal probability result called the *Borel–Cantelli lemma*, which allows upgrading convergence in probability to almost sure convergence; it just requires the slightly stronger assumption, replacing (2) above, that $\sum_N \mathrm{Var}[M^{(N)}] < \infty$. We will see that, in the ensembles we care about, with $M^{(N)} = M_k(\boldsymbol{\lambda}^{(N)})$, we'll actually have $\mathrm{Var}[M^{(N)}] = O(\frac{1}{N^2})$, and so we will actually be able to conclude almost sure convergence of the sample moments to their limit means. *Caution*: there is a technical point here, that in order to even talk about almost sure convergence, all of the random variables in question must be *defined on the same sample space*. But if we are taking larger and larger random matrices, there is no natural reason for this to be so. Therefore, it is actually a little artificial to talk about almost sure convergence in this context (although it can be technically useful).

This is what will turn out to happen to our sample moments of eigenvalue ensembles $\boldsymbol{\lambda}^{(N)}$: although they will all be random (with distributions becoming more and more complicated as $N \to \infty$), they will *concentrate about their means*, and so to compute their large-N limits, we need only compute the large-N limits of their *expected values*. Then, provided we can show also that their variances decay to 0, we will know that the random sample moments themselves converge

to deterministic values, which we can then try to identify as the true moments of some probability density function.

Let's conclude this section by introducing some slightly new notation (for normalized trace), in service of a final corollary.

DEFINITION 2.22. *Let* tr_N *denote the* **normalized trace** *on* $N \times N$ *matrices:*

$$\mathrm{tr}_N(H) = \frac{1}{N}\,\mathrm{Tr}\,(H), \qquad H \in \mathbb{M}_{N\times N}.$$

It might seem frivolous to define a whole new notation for the normalized trace, but it will turn out to be very convenient, for at least two reasons: we will use the symbol often and it will improve readability, and also the explicit N in tr_N will serve to remind us what the dimension of the underlying matrices is, which will be helpful when we start dealing with a growing N.

The following Corollary weaves together Propositions 2.15 and 2.20 with Theorem 2.13 to define precisely what our goal will be going forward.

COROLLARY 2.23. *Let* $(W_N)_{N\in\mathbb{N}}$ *be a sequence of* $N \times N$ *random matrices, with eigenvalues* $(\boldsymbol{\lambda}^{(N)})_{N\in\mathbb{N}}$*. Suppose that the following conditions hold.*

(1) *For each* $k \in \mathbb{N}$, $\lim_{N\to\infty} \mathbb{E}\,\mathrm{tr}_N[(W_N)^k] = \mu_k$ *exists.*
(2) *For each* $k \in \mathbb{N}$, $\lim_{N\to\infty} \mathrm{Var}(\,\mathrm{tr}_N[(W_N)^k]) = 0$.
(3) *There is a probability density* f*, supported on a bounded interval* $[\alpha, \beta]$*, whose moments are* μ_k*.*

Then f *is the asymptotic probability density of the ensemble of eigenvalues* $(\boldsymbol{\lambda}^{(N)})_{N\in\mathbb{N}}$*; i.e. for all* $a < b$*,*

$$\frac{1}{N}\#\{j\colon \lambda_j^{(N)} \in [a,b]\} \;\longrightarrow_{\mathbb{P}}\; \int_a^b f(\lambda)\,d\lambda \quad \textit{as} \;\; N \to \infty.$$

Note that both $\mathbb{E}\,\mathrm{tr}_N[(W_N)^k]$ and

$$\mathrm{Var}(\,\mathrm{tr}_N[(W_N)^k]) = \mathbb{E}(\,\mathrm{tr}_N[(W_N)^k]^2) - \mathbb{E}(\,\mathrm{tr}_N[(W_N)^k])^2$$

are explicitly computable for any N and k, in terms of the entries of W_N. We will now proceed to analyze these quantities for the random sample covariance matrix ensembles we are about, in the hopes of proving conditions (1)-(3) in Corollary 2.23 hold, allowing us to finally find the asymptotic probability density of the eigenvalues — i.e. to find *the shape of noise*.

2.4. Moments of Wishart Ensembles

Let's give a name to the kind of random matrix, related to sample covariance matrices of pure random noise, that are of interest to us presently.

DEFINITION 2.24. *Let* $m \geq N$ *be positive integers. A* **Wishart matrix** *of size* (m, N) *is an* $N \times N$ *matrix* $\mathbf{W} = \mathbf{W}^{(m,N)}$ *of the form* $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$*, where* $\mathbf{X} \in \mathbb{M}_{m\times N}$ *has all i.i.d. entries, with finite moments of all orders, and standardized with* $\mathbb{E}([\mathbf{X}]_{ij}) = 0$ *and* $\mathbb{E}([\mathbf{X}]_{ij}^2) = 1$*. If the entries are* $\mathcal{N}(0,1)$*, we call it a* **Gaussian Wishart matrix**.

The ensemble of eigenvalues $(\boldsymbol{\lambda}^{(N)})_{N\in\mathbb{N}}$ *of* $\mathbf{W}^{(m,N)}$ *is called a* **Wishart ensemble**.

So: the calculation in Example 2.18 showed that if $\mathbf{W}$ is a Gaussian Wishart ensemble of size (m, N), and if we send $m \to \infty$ and $N \to \infty$ in such a way that $m/N \to \varrho \in (0, \infty)$, then the sample mean $M_1(\boldsymbol{\lambda}^{(N)})$ of the associated Wishart ensemble tends to ϱ. Contemplating that

example again, note that the normal distribution of the entries played virtually no role: to get the same answer, all that was required was (1) the entries of $\mathbf{X}$ were i.i.d., and (2) the entries were standardized with mean 0 and variance 1. This is why we've made the more general definition of Wishart matrices, without the assumption that the entries are normal. We will see that, in general, the distribution of the entries doesn't matter to the shape of the asymptotic probability density of the ensemble of eigenvalues.

Following the advice of Corollary 2.23, if we want to identify the asymptotic probability density of a Wishart ensemble, we should begin by calculating the expected values of the sample moments of the ensemble, which means computing the *expected normalized traces of powers of* $\mathbf{W}$; i.e.

$$\mathbb{E}\,\mathrm{tr}_N[\mathbf{W}^k], \qquad k \in \mathbb{N}.$$

Example 2.18 did this for $k = 1$; we need to do it for all higher powers. As a warm-up, let's start with $k = 2$.

PROPOSITION 2.25. *Let* $\mathbf{W} = \mathbf{W}^{(m,N)} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$ *be a Wishart matrix of size* (m, N). *Denote the common fourth moment of the entries of* $\mathbf{X}$ *as* $\phi = \mathbb{E}([\mathbf{X}]_{ai}^4)$. *Then*

$$\mathbb{E}\,\mathrm{tr}_N(\mathbf{W}^2) = \frac{m(N+m-2)}{N^2} + \frac{m}{N^2}\phi.$$

PROOF. We begin by expanding the trace of the square of a general Wishart matrix of size (m, N):

$$\mathrm{Tr}\,(\mathbf{W}^2) = \sum_{i=1}^{N}[\mathbf{W}^2]_{ii}.$$

Now $\mathbf{W}^2 = \frac{1}{N^2}\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}$, so

$$[\mathbf{W}^2]_{ii} = \frac{1}{N^2}[\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\mathbf{X}]_{ii} = \frac{1}{N^2}\sum_{j=1}^{N}[\mathbf{X}^\top\mathbf{X}]_{ij}[\mathbf{X}^\top\mathbf{X}]_{ji}.$$

Therefore

$$\mathrm{Tr}\,(\mathbf{W}^2) = \frac{1}{N^2}\sum_{i,j=1}^{N}[\mathbf{X}^\top\mathbf{X}]_{ij}[\mathbf{X}^\top\mathbf{X}]_{ji}. \tag{2.16}$$

Now, we expand the entries $[\mathbf{X}^\top\mathbf{X}]_{ij}$ in terms of the entries of $\mathbf{X}$:

$$[\mathbf{X}^\top\mathbf{X}]_{ij} = \sum_{a=1}^{m}[\mathbf{X}^\top]_{ia}[\mathbf{X}]_{aj} = \sum_{a=1}^{m}[\mathbf{X}]_{ai}[\mathbf{X}]_{aj}.$$

This term appears twice in (2.16) (with indices reversed in the second), and so there will be two sums there; to avoid confusion, we will use a different index letter, b, for the second sum. That is,

$$\begin{aligned}\mathrm{Tr}\,(\mathbf{W}^2) &= \frac{1}{N^2}\sum_{i,j=1}^{N}\left(\sum_{a=1}^{m}[\mathbf{X}]_{ai}[\mathbf{X}]_{aj}\right)\left(\sum_{b=1}^{m}[\mathbf{X}]_{bj}[\mathbf{X}]_{bi}\right)\\ &= \frac{1}{N^2}\sum_{i,j=1}^{N}\sum_{a,b=1}^{m}[\mathbf{X}]_{ai}[\mathbf{X}]_{aj}[\mathbf{X}]_{bi}[\mathbf{X}]_{bj}.\end{aligned}$$

Hence, the desired quantity (recalling that $\operatorname{tr}_N = \frac{1}{N}\operatorname{Tr}$, and using the linearity of expectation) is

$$\mathbb{E}\operatorname{tr}_N(\mathbf{W}^2) = \frac{1}{N^3}\sum_{i,j=1}^{N}\sum_{a,b=1}^{m}\mathbb{E}\left([\mathbf{X}]_{ai}[\mathbf{X}]_{aj}[\mathbf{X}]_{bi}[\mathbf{X}]_{bj}\right). \tag{2.17}$$

We must now compute the expected values in the terms of this sum. The key observation to make this possible is that the entries $\{[\mathbf{X}]_{ai} \colon 1 \le a \le m, 1 \le i \le N\}$ are all independent, and all have mean 0. For example: suppose $(a, b, i, j) = (1, 2, 1, 4)$; then

$$\mathbb{E}([\mathbf{X}]_{ai}[\mathbf{X}]_{aj}[\mathbf{X}]_{bi}[\mathbf{X}]_{bj}) = \mathbb{E}([\mathbf{X}]_{11}[\mathbf{X}]_{14}[\mathbf{X}]_{21}[\mathbf{X}]_{24}).$$

All four entries are thus independent. In particular, the first entry is independent of the last three, and so we can factor the expectation:

$$\mathbb{E}([\mathbf{X}]_{11}[\mathbf{X}]_{14}[\mathbf{X}]_{21}[\mathbf{X}]_{24}) = \mathbb{E}([\mathbf{X}]_{11}) \cdot \mathbb{E}([\mathbf{X}]_{14}[\mathbf{X}]_{21}[\mathbf{X}]_{24}) = 0 \cdot (\text{something}) = 0.$$

What we see from this example is the following: most of the terms in (2.17) are 0. More precisely: in any of the N^2m^2 terms where there is at least one index pair (a, i) that doesn't exactly match any of the other three index pairs, the resulting expected value is 0.

Therefore, to find terms that do contribute, we need to look for indices (i, j, a, b) that leave none of the pairs "lonely". We can do this systematically as follows. Consider the first index (a, i); it must match (at least) one of the other three. We take them in turns.

- *The first two entries are equal:* $[\mathbf{X}]_{ai} = [\mathbf{X}]_{aj}$. This means $i = j$. This leave the last two entries $[\mathbf{X}]_{bi}[\mathbf{X}]_{bj}$; but since $i = j$, these two are also equal, so no one is lonely: these terms look like $\mathbb{E}([\mathbf{X}]_{ai}^2[\mathbf{X}]_{bi}^2)$.
- *The first and third entries are equal:* $[\mathbf{X}]_{ai} = [\mathbf{X}]_{bi}$. This means $a = b$. Now considering the remaining entries $[\mathbf{X}]_{aj}[\mathbf{X}]_{bj}$, we see they are also equal, so no one is lonely. These terms look like $\mathbb{E}([\mathbf{X}]_{ai}^2[\mathbf{X}]_{aj}^2)$.
- *The first and fourth entries are equal:* $[\mathbf{X}]_{ai} = [\mathbf{X}]_{bj}$. This means $a = b$ and $i = j$. In this case, all four of the entries are equal, and we have $\mathbb{E}([\mathbf{X}]_{ai}^4)$.

Hence, the only terms that contribute to the sum in (2.17) are those where either $i = j$ or $a = b$ (or perhaps both). There is some overlap in these, so we separate them carefully:

- $i = j, a \neq b$: there are $Nm(m-1)$ of these terms, and they are all equal to $\mathbb{E}([\mathbf{X}]_{ai}^2[\mathbf{X}]_{bi}^2) = \mathbb{E}([\mathbf{X}]_{ai}^2)\mathbb{E}([\mathbf{X}]_{bi}^2) = 1 \cdot 1$ by the independence and the assumption that the second moment of each entry is 1.
- $i \neq j, a = b$: there are $N(N-1)m$ of these terms, and they are all equal to $\mathbb{E}([\mathbf{X}]_{ai}^2[\mathbf{X}]_{aj}^2) = \mathbb{E}([\mathbf{X}]_{ai}^2)\mathbb{E}([\mathbf{X}]_{aj}^2) = 1 \cdot 1 = 1$ as above.
- $i = j, a = b$: there are Nm of these terms, and they are all equal to $\mathbb{E}([\mathbf{X}]_{ai}^4) = \mathbb{E}([\mathbf{X}]_{11}^4)$, the common fourth moment of all the entries.

Thus, the sum in (2.17) is equal to

$$\begin{aligned}\mathbb{E}\operatorname{tr}_N(\mathbf{W}^2) &= \frac{1}{N^3}\left(Nm(m-1) + N(N-1)m + Nm\mathbb{E}([\mathbf{X}]_{11}^4)\right) \\ &= \frac{m(N+m-2)}{N^2} + \frac{m}{N^2}\phi. \end{aligned} \tag{2.18}$$

□

The exact quantity above depends on the distributions of the entries of the "noise" matrix $\mathbf{X}$: it depends on the fourth moment ϕ of the entries. It also depends explicitly on both m and N. Comparing to Example 2.18, we saw there that $\mathbb{E}\operatorname{tr}_N(\mathbf{W}) = \frac{m}{N}$ depended only on the ratio $\frac{m}{N}$, and

not on the distribution of the entries. Nevertheless, in the present setting, the same can be said in the large-N limit. Let's make this precise.

COROLLARY 2.26. *Fix $\varrho > 0$. For each $N \in \mathbb{N}$, let m_N be a sequence of positive integers with the property that $\lim_{N\to\infty} \frac{m_N}{N} = \varrho$. Let $\mathbf{W} = \mathbf{W}^{(m_N,N)}$ be a Wishart matrix of size (m_N, N). Then*

$$\lim_{N\to\infty} \mathbb{E}\,\mathrm{tr}_N(\mathbf{W}^2) = \varrho^2 + \varrho.$$

PROOF. From Proposition 2.25, we have

$$\mathbb{E}\,\mathrm{tr}_N(\mathbf{W}^2) = \frac{m_N(N + m_N - 2)}{N^2} + \frac{m_N}{N^2}\phi = \frac{m_N}{N} + \frac{m_N^2}{N^2} + \frac{(\phi - 2)m_N}{N^2}.$$

The first two terms converge to $\varrho + \varrho^2$, while the third is $\frac{\phi-2}{N} \cdot \frac{m_N}{N}$ and so converges to 0. □

We now have a clear indication that the correct scaling regime to find a limit is to let $m, N \to \infty$ simultaneously, with aspect ratio $\frac{m}{N}$ approaching some limit ϱ. We'll now attempt to mimic the proof of Proposition 2.25 to calculate the general kth moment. To begin, we use induction to expand the trace of a power of $\mathbf{W}$:

$$\mathrm{Tr}\,(\mathbf{W}^k) = \sum_{i_1=1}^{N} [\mathbf{W}^k]_{i_1 i_1} = \sum_{i_1,i_2=1}^{N} [\mathbf{W}]_{i_1 i_2}[\mathbf{W}^{k-1}]_{i_2 i_1} = \sum_{i_1,i_2,i_3=1}^{N} [\mathbf{W}]_{i_1 i_2}[\mathbf{W}]_{i_2 i_3}[\mathbf{W}^{k-2}]_{i_2 i_1}$$

Continuing this way, we see that

$$\mathrm{Tr}\,(\mathbf{W}^k) = \sum_{i_1,\dots,i_k=1}^{N} [\mathbf{W}]_{i_1 i_2}[\mathbf{W}]_{i_2 i_3} \cdots [\mathbf{W}]_{i_{k-1} i_k}[\mathbf{W}]_{i_k i_1}. \tag{2.19}$$

Now, $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$, and so

$$[\mathbf{W}]_{ij} = \frac{1}{N}\sum_{a=1}^{m} [\mathbf{X}^\top]_{ia}[\mathbf{X}]_{aj} = \frac{1}{N}\sum_{a=1}^{m} [\mathbf{X}]_{ai}[\mathbf{X}]_{aj}.$$

Applying this term-by-term in (2.19), each of the k terms gives rise to a new sum, and incorporating the powers of N, we have

$$\mathrm{tr}_N(\mathbf{W}^k) = \frac{1}{N^{k+1}} \sum_{i_1,\dots,i_k=1}^{N} \sum_{a_1,a_2,\dots,a_k=1}^{m} [\mathbf{X}]_{a_1 i_1}[\mathbf{X}]_{a_1 i_2}[\mathbf{X}]_{a_2 i_2}[\mathbf{X}]_{a_2 i_3} \cdots [\mathbf{X}]_{a_k i_k}[\mathbf{X}]_{a_k i_1}. \tag{2.20}$$

On its face, the scaling here does not look right. The sum (2.20) has $N^k m^k$ terms, while the scaling is $\frac{1}{N^k}$; if all the terms were the same size, we would have $\mathrm{tr}_N(\mathbf{W}^k) \sim \frac{m^k}{N}$, and as we are expecting the right scaling regime to be $\frac{m}{N} \sim \varrho$ constant, this would result in blow-up as $m \to \infty$. So, in order for this to make sense, we must see that the *vast* majority of the terms in (2.20) are "small".

We are first interested in $\mathbb{E}\,\mathrm{tr}_N(\mathbf{W}^k)$, and so taking expectations, we must understand the expectation of the general term:

$$\mathbb{E}\left([\mathbf{X}]_{a_1 i_1}[\mathbf{X}]_{a_1 i_2}[\mathbf{X}]_{a_2 i_2}[\mathbf{X}]_{a_2 i_3} \cdots [\mathbf{X}]_{a_k i_k}[\mathbf{X}]_{a_k i_1}\right). \tag{2.21}$$

The independence and identical distribution of the distinct entries $[\mathbf{X}]_{ai}$ implies that this quantity doesn't depend very much on what the actual indices $i_1, \dots, i_k$ and $a_1, \dots, a_k$ are; rather, all that matters is how they are grouped together according to being identical or not. Indeed,

$$\mathbb{E}\left([\mathbf{X}]_{13}[\mathbf{X}]_{14}[\mathbf{X}]_{13}[\mathbf{X}]_{21}\right) = \mathbb{E}\left([\mathbf{X}]_{27}[\mathbf{X}]_{54}[\mathbf{X}]_{27}[\mathbf{X}]_{98}\right) \tag{2.22}$$

because, after grouping according to which index-pairs match, we have

$$\mathbb{E}\left([\mathbf{X}]_{13}[\mathbf{X}]_{14}[\mathbf{X}]_{13}[\mathbf{X}]_{21}\right) = \mathbb{E}\left([\mathbf{X}]_{13}^2[\mathbf{X}]_{14}[\mathbf{X}]_{21}\right) = \mathbb{E}\left([\mathbf{X}]_{13}^2\right)\mathbb{E}\left([\mathbf{X}]_{14}\right)\mathbb{E}\left([\mathbf{X}]_{21}\right)$$
$$\mathbb{E}\left([\mathbf{X}]_{27}[\mathbf{X}]_{54}[\mathbf{X}]_{27}[\mathbf{X}]_{98}\right) = \mathbb{E}\left([\mathbf{X}]_{27}^2[\mathbf{X}]_{27}[\mathbf{X}]_{98}\right) = \mathbb{E}\left([\mathbf{X}]_{27}^2\right)\mathbb{E}\left([\mathbf{X}]_{27}\right)\mathbb{E}\left([\mathbf{X}]_{98}\right)$$

and both of these equal

$$\mathbb{E}\left([\mathbf{X}]_{11}^2\right)\mathbb{E}\left([\mathbf{X}]_{11}\right)^2$$

because the entries all have the same distribution.

Thus, in a general term like (2.21), we only need to keep track of the **partition** induced by the indices.

DEFINITION 2.27. *A* **partition** *of a finite set A is a collection of non-empty disjoint subsets of A whose union is all of A. We denote partitions usually by $\pi = \{B_1, \ldots, B_r\}$ where the subsets B_ℓ are called the* **blocks** *of the partition. For example: $\pi = \{\{1,3,4\},\{2\},\{5,6\}\}$ is a partition of the set $[6] = \{1,2,3,4,5,6\}$.*

Each expression (2.21) gives rise to a partition of the $2k$ terms in the product: they are partitioned according to their indices (a, i), i.e. grouped according to which terms have identical indices. This is the only information that determines the value of the term. That is, two expressions

$$\mathbb{E}\left([\mathbf{X}]_{a_1 i_1}[\mathbf{X}]_{a_1 i_2}[\mathbf{X}]_{a_2 i_2}[\mathbf{X}]_{a_2 i_3}\cdots[\mathbf{X}]_{a_k i_k}[\mathbf{X}]_{a_k i_1}\right),$$
$$\mathbb{E}\left([\mathbf{X}]_{b_1 j_1}[\mathbf{X}]_{b_1 j_2}[\mathbf{X}]_{b_2 j_2}[\mathbf{X}]_{b_2 j_3}\cdots[\mathbf{X}]_{b_k j_k}[\mathbf{X}]_{b_k j_1}\right)$$

are, in fact, equal, provided that their indices both have the same partition $\pi = \{B_1, \ldots, B_r\}$, where each block B_s is a collection of matrix index labels all assigned the same numbers within the block. For example: in the two (equal) terms in (2.22), the common partition is $\{\{1,3\},\{2\},\{4\}\}$, indicating that the first and third terms are equal, and distinct from the two other non-equal second and fourth terms. As the following discussion shows, by the independence and identical distribution of the entries, the value of the expectation (2.21) depends only on the sizes and number of blocks in π: (2.21) is equal to

$$\mathbb{E}_\pi := \mathbb{E}([\mathbf{X}]_{11}^{\#B_1})\mathbb{E}([\mathbf{X}]_{11}^{\#B_2})\cdots\mathbb{E}([\mathbf{X}]_{11}^{\#B_r}). \tag{2.23}$$

Summarizing the above discussion, we have the following.

LEMMA 2.28. *Given indices $\mathbf{a} = (a_1, \ldots, a_k)$ and $\mathbf{i} = (i_1, \ldots, i_k)$, let $\pi(\mathbf{a}, \mathbf{i})$ denote the partition of those indices in the term $\mathbb{E}\left([\mathbf{X}]_{a_1 i_1}[\mathbf{X}]_{a_1 i_2}[\mathbf{X}]_{a_2 i_2}[\mathbf{X}]_{a_2 i_3}\cdots[\mathbf{X}]_{a_k i_k}[\mathbf{X}]_{a_k i_1}\right)$. Let $\mathcal{P}(2k)$ denote the set of all partitions of $[2k]$. Then*

$$\mathbb{E}\,\mathrm{tr}_N(\mathbf{W}^k) = \frac{1}{N^{k+1}} \sum_{\pi\in\mathcal{P}(2k)} \sum_{\substack{\mathbf{a}\in[m]^k, \mathbf{i}\in[N]^k \\ \pi(\mathbf{a},\mathbf{i})=\pi}} \mathbb{E}_\pi. \tag{2.24}$$

This key observation now shows almost immediately why most of the terms in (2.24) are small. For a given term (2.21), suppose there is at least one block B_s in its partition π that is a singleton, i.e. $\#B_s = 1$. That means the term is equal to $\mathbb{E}([X]_{11}^1) \cdot \text{stuff} = 0$ because all the terms are centered. In other words:

LEMMA 2.29. *If $\pi \in \mathcal{P}(2k)$ contains a singleton block, then $\mathbb{E}_\pi = 0$.*

This explains the apparent scaling problem. Hence, only partitions in which every block has at least two elements contribute to the sum (2.24) in the large-N limit. Let's consider two example partitions where *all* blocks have exactly two elements.

EXAMPLE 2.30. Suppose the partition in question is $\sigma = \{\{1,2\},\{3,4\},\dots,\{2k-1,2k\}\}$. Examining the corresponding indices for terms at those places in (2.21), this partition would mean that

$$(a_1,i_1) = (a_1,i_2), \quad (a_2,i_2) = (a_2,i_3), \quad \dots \quad (a_k,i_k) = (a_k,i_1).$$

The first coordinates in each pair of terms already automatically, but the incidence of the second ones mean that $i_1 = i_2 = \cdots = i_k$. Hence, the number of $(\mathbf{a},\mathbf{i})$ for which $\pi(\mathbf{a},\mathbf{i}) = \sigma$ is $m^k N$. Since each block in σ has size 2, $\mathbb{E}_\sigma = \mathbb{E}([\mathbf{X}]_{11}^2)^k = 1$, and so these terms contribute

$$\frac{1}{N^{k+1}} \sum_{\substack{\mathbf{a}\in[m]^k,\mathbf{i}\in[N]^k \\ \pi(\mathbf{a},\mathbf{i})=\sigma}} \mathbb{E}_\sigma = \frac{1}{N^{k+1}} \cdot m^k N \cdot 1 = \frac{m^k}{N^k}$$

to $\mathbb{E}\,\mathrm{tr}_N(\mathbf{W}^k)$, and this (by assumption) has a limit (ϱ^k) as $N, m \to \infty$.

EXAMPLE 2.31. For a related example, suppose that the partition of indices is

$$\{\{1,3\},\{2,4\},\{5,6\},\{7,8\}\dots,\{2k-1,2k\}\}.$$

This is almost the same as the example we considered above, but now the first two blocks in π have been "twisted up" to cross each other. Examining the indices in (2.21), such terms satisfy

$$(a_1,i_1) = (a_2,i_3), \quad (a_1,i_2) = (a_2,i_2), \quad (a_3,i_3) = (a_3,i_4), \quad \dots \quad (a_k,i_k) = (a_k,i_1).$$

The first two equations say $a_1 = a_2$ and $i_2 = i_3$; all the remaining specify that $i_3 = i_4 = i_5 = \cdots = i_k = i_1$. Hence, we still have $i_1 = i_2 = \cdots = i_k$, as above, but here we also have the constraint that $a_1 = a_2$. So the total number of terms in the inner sum in (2.24)is $m^{k-1}N$. Combining this with the overall $\frac{1}{N^{k+1}}$ scaling, and noting that each such term (with this as its index partition) has the value $\mathbb{E}_\pi = 1$, means that these terms contribute $\frac{m^{k-1}}{N^k}$ to the sum; in the scaling regime where $\frac{m}{N} \to \varrho$, $\frac{m^{k-1}}{N^k} \to 0$ as $m, N \to \infty$. Hence, these terms don't contribute in the limit.

The task from here is a little complicated. We have to consider each possible partition $\pi \in \mathcal{P}_{\geq 2}(2k)$ (the partitions in which every block has at least two elements), and compute both $\mathbb{E}_\pi$ (this is easy, from (2.23) it is a product of moments of the entries) and the number of indices $(\mathbf{a},\mathbf{i}) \in [m]^k \times [N]^k$ that give rise to that partition. It's a beautiful story, involving several different kinds of interesting combinatorial structures (non-crossing partitions, lattice path, directed graphs) that go beyond what we have time to study presently. Here is a brief summary of how it turns out:

- If π contains *any* blocks with *more* than 3 elements, the number of terms in the inner sum in (2.24) corresponding to that partition π is $o(N^{k+1})$, hence do not contribute in the limit.
- Together with the observation that all "lonely" partitions π (containing a singleton block) have $\mathbb{E}_\pi = 0$, this means that, in the large-N limit, the sum is restricted to $\pi \in \mathcal{P}_2(2k)$: the partitions in which each block has exactly two elements. These are called *pairings* or *matchings*. Note: from (2.23), if π is any paring, $\mathbb{E}_\pi = \mathbb{E}([\mathbf{X}]_{11}^2)^k = 1$. Thus, we see that no higher moments of the distribution of the matrix entries play a role in the large-N limit. This is the simplest example of a *universality* result in random matrix theory: the asymptotic behavior is universal, regardless of the distribution of the entries.
- Not *all* pairings contribute in the large-N limit, as Example 2.31 shows. Determining exactly which ones contribute to leading order, and what the contribution is (as a function of $\frac{m}{N}$) is a dicey proposition that requires translation to other combinatorial objects we know how to count. The interested reader can consult my Random Matrix Theory notes, also

posted on the course webpage, in particular Section 4.1, for a taste of the combinatorics involved in a related but slightly simpler ensemble.

- Similar, but somewhat more involved, combinatorial arguments also show that

$$\operatorname{Var}\left(\operatorname{tr}_N(\mathbf{W}^k)\right) = O\left(\frac{1}{N^2}\right)$$

in the scaling regime $\frac{m}{N} \to \varrho$, and so we get the desired concentration about the mean.

Let us now record the final result that comes from all this.

THEOREM 2.32. *Fix $\varrho > 0$. For each $N \in \mathbb{N}$, let m_N be a sequence of positive integers with the property that $\lim_{N\to\infty} \frac{m_N}{N} = \varrho$. Let $\mathbf{W} = \mathbf{W}^{m_N,N}$ be a Wishart matrix of size $m_N \times N$. Then for integers $k \geq 1$,*

$$\lim_{N\to\infty} \operatorname{tr}_N(\mathbf{W}^k) = \sum_{r=1}^{k} \frac{1}{r}\binom{k}{r-1}\binom{k-1}{r-1}\varrho^r.$$

where the limit is in probability.

We conclude this section with some remarks about our choice to work with $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$, rather than the actual sample covariance matrix $\mathbf{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top$. This choice was made for efficiency, since we are assuming $m \geq N$: $\mathbf{W}$ is $N \times N$ while $\mathbf{C}$ is $m \times m$. The two matrices have the same non-zero eigenvalues (as you proved on Homework 1), and so the only difference in the histograms of eigenvalues will be the presence (or lack) of a spike of eigenvalues at 0.

Let's explore how that choice will affect the moments. The relation is quite simple, but we have to be careful. Note that

$$\operatorname{Tr}(\mathbf{C}^k) = \operatorname{Tr}\left[\left(\frac{1}{N}\mathbf{X}\mathbf{X}^\top\right)^k\right] = \frac{1}{N^k}\operatorname{Tr}(\mathbf{X}\mathbf{X}^\top\cdots\mathbf{X}\mathbf{X}^\top).$$

Using the cyclic invariance of the trace (Lemma 2.16), we can move one $\mathbf{X}^\top$ from the end to the beginning, so

$$\operatorname{Tr}(\mathbf{C}^k) = \frac{1}{N^k}\operatorname{Tr}(\mathbf{X}^\top\mathbf{X}\cdots\mathbf{X}^\top\mathbf{X}) = \operatorname{Tr}(\mathbf{W}^k).$$

It seems the moments are the same; but the subtlety is that we were computing the *normalized* trace (i.e. the sample moments of the eigenvalue ensemble, which involves dividing by the number of eigenvalues). Since $\mathbf{C}$ is $m \times m$, this means

$$\operatorname{tr}_m(\mathbf{C}^k) = \frac{1}{m}\operatorname{Tr}(\mathbf{C}^k) = \frac{1}{m}\operatorname{Tr}(\mathbf{W}^k) = \frac{N}{m}\operatorname{tr}_N(\mathbf{W}^k).$$

In the scaling regime $\frac{m}{N} \to \varrho$, this is about $\varrho^{-1}\operatorname{tr}_N(\mathbf{W}^k)$, which means we have the following corollary.

COROLLARY 2.33. *Fix $\varrho > 0$. For each $N \in \mathbb{N}$, let m_N be a sequence of positive integers with the property that $\lim_{N\to\infty} \frac{m_N}{N} = \varrho$. Let $\mathbf{X}$ be an $m_N \times N$ feature matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and let $\mathbf{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top$ be its sample covariance matrix . Then for integers $k \geq 1$,*

$$\lim_{N\to\infty} \operatorname{tr}_N(\mathbf{W}^k) = \sum_{r=1}^{k} \frac{1}{r}\binom{k}{r-1}\binom{k-1}{r-1}\varrho^{r-1} = \sum_{r=0}^{k-1} \frac{1}{r+1}\binom{k}{r}\binom{k-1}{r}\varrho^r.$$

where the limit is in probability.

REMARK 2.34. One can think of the relationship between $\mathbf{W}$ and $\mathbf{C}$ as an exchange $m \leftrightarrow N$: if we let $\mathbf{Y} = \mathbf{X}^\top$, then $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^\top$. But this is not quite the right thing to do, since the new "feature matrix" $\mathbf{Y}$ is $N \times m$, which means that there are m columns, not N. Hence, the covariance matrix associated to $\mathbf{Y}$ is $\frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$. The relationship is then

$$\mathbf{C}_\mathbf{Y} = \frac{1}{m}\mathbf{Y}\mathbf{Y}^\top = \frac{N}{m}\cdot\frac{1}{N}\mathbf{Y}\mathbf{Y}^\top = \frac{N}{m}\mathbf{W}.$$

This new covariance matrix is still $N \times N$, but the scaling of the matrix itself has changed (by $\frac{N}{m} \cong \varrho$), and that means that the kth moment will change by a factor of ϱ^{-k}, not just ϱ^{-1}. We will explore these various exchanges and rescalings in Section 2.6.

2.5. Moment Generating Function(s)

We have now completed steps (1) and (2) in the plan set out at the end of Section 2.2: we have computed the large-N limits μ_k of the sample moments M_k of the eigenvalue ensembles $\boldsymbol{\lambda}^{(N)}$ for our sample covariance matrices. The final step in the plan is to use these moments to determine the asymptotic density of these ensembles: i.e. the limit density underlying the histograms in, for example, Figure 2.1.

The general question before us, then, is: given the moments of a distribution, how do we reconstruct the distribution itself? In principle we could follow the procedure in the proof of Theorem 2.13, but already in the limit: if we knew already that our desired density f were supported inside some interval $[\alpha, \beta]$, we could compute $\int_\alpha^t f(x)\,dx$ using the Bernstein polynomials $B_n^{(t)}$ for the indicator function $\mathbb{1}_{[\alpha,t]}$:

$$\int_\alpha^t f(x)\,dx = \int_\alpha^\beta \mathbb{1}_{[\alpha,t]}(x)f(x)\,dx = \int_\alpha^\beta \lim_{n\to\infty} B_n^{(t)}(x)f(x)\,dx.$$

We can then explicitly write the Bernstein polynomials in terms of their coefficients $B_n^{(t)}(x) = \sum_{k=0}^n b_{n,k}^{(t)} x^k f(x)\,dx$, and (interchanging the limit and the integral) find

$$\int_\alpha^t f(x)\,dx = \lim_{n\to\infty}\int_\alpha^\beta \sum_{k=0}^n b_{n,k}^{(t)}x^k f(x)\,dx = \lim_{n\to\infty}\sum_{k=0}^n b_{n,k}^{(t)}\int_\alpha^\beta x^k f(x)\,dx.$$

All we know about f so far is that this inside integral is μ_k; thus, if we knew f to be continuous (so the Fundamental Theorem of Calculus would apply), we would reconstruct f by differentiating

$$f(t) = \frac{d}{dt}\lim_{n\to\infty}\sum_{k=0}^n b_{n,k}^{(t)}\,\mu_k.$$

The coefficients $b_{n,k}^{(t)}$ are fairly explicit, from the proof of Bernstein's Lemma 2.7, so we might hope to actually do this computation. There are a few problems with it, though: first, we don't know a priori what enveloping $[\alpha, \beta]$ to use (or if there even is one), and this heavily influences the formula for the $b_{n,k}^{(t)}$ coefficients. Also: we have no a priori reason to think the limit density is continuous. More subtly: we might want to interchange the derivative and the limit above, but that is *really* problematic (even if the convergence is uniform, such an interchange often fails).

We therefore want a (computationally) better way to recover the density from its known moments. One possibility is the **moment-generating function**. If Y is a random variable, its MGF is

$$M_Y(t) = \mathbb{E}[e^{tY}] = \int_\mathbb{R} e^{ty} f_Y(y)\,dy$$

(where the second equality holds if Y has a PDF f_Y). This function doesn't always make sense: the random variable e^{tY} need not have a finite expectation for any particular t (other than $t = 0$). It is a theorem, though, that if $M_Y(t)$ is finite for t in some open interval, then that function uniquely determines the distribution of Y. This can be understood as a condition on moments: expanding the exponential power series, we have

$$M_Y(t) = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{t^k}{k!} Y^k\right] = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}(Y^k). \tag{2.25}$$

The second equality requires justification and does not always hold true; we could, for example, use Fubini's theorem to justify the $\mathbb{E}$-sum interchange, provided we know that the right-hand-side converges (uniformly enough). From the root test, we know this power series will converge on some neighborhood of 0, provided that

$$\lim_{k\to\infty} \left(\frac{1}{k!}\mathbb{E}(|Y|^k)\right)^{1/k} < \infty.$$

That is, provided there is some constant $C > 0$ so that $\mathbb{E}(|Y|^k) \leq C^k k!$, the series defining $M_Y(t)$ will converge (on at least the interval $|t| < 1/C$), and hence determined the distribution of Y. The function M_Y is completely determined by the moments of Y (cf. (2.25)), and so this ought to provide a way to determine the distribution from the moments.

But *how* do we directly recover f_Y from M_Y? The best answer requires we turn our head $90°$ to the side, and allow t to be an *imaginary* number.

DEFINITION 2.35. *Let Y be a random variable. Its* **characteristic function** *χ_Y is the complex-valued function on $\mathbb{R}$ defined by*

$$\chi_Y(t) = \mathbb{E}(e^{itY}).$$

If Y has a probability density function f_Y, then $\chi_Y(t) = \int_{\mathbb{R}} e^{ity} f_Y(y)\, dy$. This is also called the **Fourier transform** *of f_Y, and sometimes denoted $\hat{f}_Y(t)$.*

One nice feature of the characteristic function is that it is *always* finite, no matter what the distribution of Y is (or if it has any finite moments at all). This is because Y is real-valued, and so $e^{itY} = \cos(tY) + i\sin(tY)$ is a complex number of unit modulus for all t. Hence, the expected value cannot blow up. Put another way: if Y has a density, then the integrals

$$a(t) = \int_{\mathbb{R}} \cos(ty) f_Y(y)\, dy \qquad \text{and} \qquad b(t) = \int_{\mathbb{R}} \sin(ty) f_Y(y)\, dy$$

are definitely both finite, since f_Y is a positive integrable function, and so $\chi_Y(t) = a(t) + ib(t)$ is well-defined. *If Y* has all finite moments, then $\chi_Y(t)$ can be written as a power-series in terms of them (just replace t with it in (2.25). More importantly, it is possible to recover the density (when it exists) from the characteristic function, as follows.

THEOREM 2.36 (Fourier Inversion). *Let f be a probability density, and let $\hat{f}$ be its Fourier transform. Then, for almost every $y \in \mathbb{R}$,*

$$f(y) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ity} \hat{f}(t)\, dt.$$

That is: you can undo the Fourier transform by taking the Fourier transform again, just with a negative sign thrown in to the argument (and an extra normalizing factor of 2π).

We will not prove Theorem 2.36. For one, it is a bit technical and outside our toolbox. For another, although it would be in principle possible to use it to find the density of eigenvalues for the Wishart ensemble, in practice there is a different tool we will use that is better adapted to the task. It is called the **Stieltjes transform**. Like the Fourier transform, it requires a complex argument.

DEFINITION 2.37. *Let Y be a random variable. Let $\mathbb{C}_+ = \{x + iy \colon y > 0\}$ denote the upper half-plane. The* **Stieltjes transform** *of Y is the function $G_Y \colon \mathbb{C}_+ \to \mathbb{C}$ defined by*

$$G_Y(z) = \mathbb{E}\left(\frac{1}{z - Y}\right).$$

If Y has a probability density f_Y, then

$$G_Y(z) = \int_{\mathbb{R}} \frac{f_Y(t)}{z - t}\, dt.$$

Note: for any fixed $\epsilon > 0$, if $z = x + iy$ with $y > \epsilon$, then $|z - Y|^2 = |x - Y + iy|^2 = (x - Y)^2 + y^2 > \epsilon^2$, which means that $|\frac{1}{z-Y}| \leq \frac{1}{\epsilon}$. Hence,

$$\text{for } \operatorname{Im}(z) > \epsilon, \quad |G_Y(z)| \leq \mathbb{E}\left|\frac{1}{z - Y}\right| \leq \mathbb{E}\left(\frac{1}{\epsilon}\right) = \frac{1}{\epsilon} < \infty.$$

Since every point $z \in \mathbb{C}_+$ has $\operatorname{Im}(z) > \epsilon$ for some $\epsilon > 0$, we see that $G_Y(z)$ converges everywhere on $\mathbb{C}_+$. It might well blow up if $z \in \mathbb{R}$, so we'll stay off the real line.

REMARK 2.38. The same arguments apply to $z \in \mathbb{C}_-$, the *lower* half-plane. Also, the interested reader might like to compute that G_Y maps $\mathbb{C}_+$ onto $\mathbb{C}_-$, and vice versa.

EXAMPLE 2.39. Suppose U is a $\operatorname{Uniform}[0, 1]$ random variable. Then

$$G_U(z) = \int_{\mathbb{R}} \frac{f_U(t)}{z - t}\, dt = \int_0^1 \frac{dt}{z - t}.$$

If z were real (and outside $[0, 1]$), we would compute this integral as $\ln(z) - \ln(z - 1) = \ln \frac{z}{z-1}$. This is also the correct answer for $z \in \mathbb{C}_+$, provided we are comfortable taking logarithms of complex numbers. For an alternative expression, let's "realize" the denominator, written in terms of $z = x + iy$:

$$\frac{1}{z - t} = \frac{\overline{z - t}}{|z - t|^2} = \frac{x - iy - t}{(x - t)^2 + y^2}$$

and so

$$G_U(x + iy) = \int_0^1 \frac{x - t}{(x - t)^2 + y^2}\, dt - iy \int_0^1 \frac{dt}{(x - t)^2 + y^2}.$$

We can evaluate both of these integrals explicitly:

$$G_U(x + iy) = \frac{1}{2} \ln\left(\frac{x^2 + y^2}{(x - 1)^2 + y^2}\right) - i\left(\arctan\left(\frac{x}{y}\right) - \arctan\left(\frac{x - 1}{y}\right)\right).$$

What can we learn from this function? Let's focus on the (negative) imaginary part

$$-\operatorname{Im} G_U(x + iy) = \left(\arctan\left(\frac{x}{y}\right) - \arctan\left(\frac{x - 1}{y}\right)\right).$$

We've only computed this when $y > 0$, but what happens as $y \downarrow 0$? The answer depends on x. The reader should verify that

$$\lim_{y \downarrow 0} -\operatorname{Im} G_U(x + iy) = \pi \mathbb{1}_{[0,1]}(x).$$

But that's (π times) the original density of U!

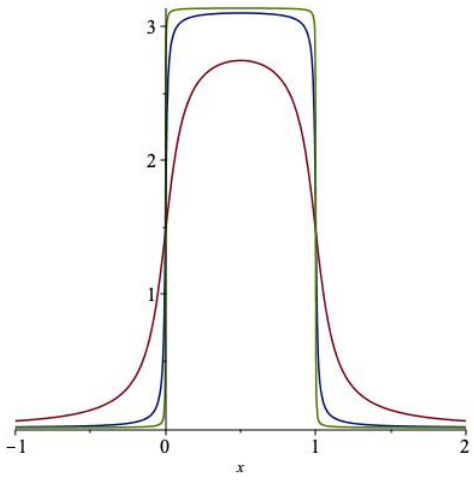


FIGURE 2.3. The graph of the negative imaginary part of the Steiltjes transform $G_U(x+iy)$ with $y = 0.1$, $y = 0.01$, and $y = 0.001$.

The result of Example 2.39 is not an accident. It works in general, and provides a robust way to recover a probability density from its Stieltjes transform. We state this precisely as follows, in the special case that the distribution is given by a continuous probability density function.

THEOREM 2.40 (Stieltjes Inversion). *Let f be a continuous probability density function, and let $G\colon \mathbb{C}_+ \to \mathbb{C}_-$ be its Stieltjes transform:*

$$G(z) = \int_{\mathbb{R}} \frac{f(t)}{z-t}\,dt.$$

Then for all $x \in \mathbb{R}$,

$$-\frac{1}{\pi}\lim_{y\downarrow 0} G(x+iy) = f(x).$$

PROOF. Following the calculations in Example 2.39, but with a general density, we have

$$G(x+iy) = \int_{\mathbb{R}} \frac{f(t)(x-t-iy)}{(x-t)^2+y^2}\,dt$$

and so

$$-\operatorname{Im} G(x+iy) = \int_{\mathbb{R}} \frac{y}{(x-t)^2+y^2} f(t)\,dt.$$

Making the change of variables $s = t - x$ yields

$$-\operatorname{Im} G(x+iy) = \int_{\mathbb{R}} \frac{y}{s^2+y^2} f(x+s)\,ds.$$

Taking f out of the picture for the moment, note that (as computed above) for any $y > 0$

$$\int_{\mathbb{R}} \frac{y}{s^2+y^2}\,ds = \arctan\left(\frac{s}{y}\right)\Big|_{s=-\infty}^{\infty} = \pi.$$

In other words: $\varphi_y(s) = \frac{1}{\pi}\frac{y}{s^2+y^2}$ is a probability density (it is evidently strictly positive). Indeed, this is a rescaling of the *Cauchy density* (the usual example of a probability density with no finite moments).

We now use a trick: since φ_y is a probability density, for any constant c, $\int_{\mathbb{R}} c\varphi_y(s)\,ds = c$. We apply this with $c = f(x)$, and compute that

$$\begin{aligned} -\frac{1}{\pi}\operatorname{Im} G(x+iy) - f(x) &= \int_{\mathbb{R}} \varphi_y(s) f(x+s)\,ds - \int_{\mathbb{R}} f(x)\varphi_y(s)\,ds \\ &= \int_{\mathbb{R}} [f(x+s) - f(x)]\varphi_y(s)\,ds. \end{aligned} \tag{2.26}$$

Now, fix any small $\epsilon > 0$. Because f is continuous at x, for all sufficiently small s, say $|s| < \delta$, we have $|f(x+s) - f(x)| < \epsilon$. Let's break up the integral as such:

$$\begin{aligned} &\left| \int_{\mathbb{R}} [f(x+s) - f(x)]\varphi_y(s)\,ds \right| \\ &\leq \int_{\mathbb{R}} |f(x+s) - f(x)|\varphi_y(s)\,ds \\ &= \int_{-\delta}^{\delta} |f(x+s) - f(x)|\varphi_y(s)\,ds + \int_{|s|\geq\delta} |f(x+s) - f(x)|\varphi_y(s)\,ds. \end{aligned}$$

In the first integral, where $|s| < \delta$, by assumption $|f(x+s) - f(x)| < \epsilon$, and so

$$\int_{-\delta}^{\delta} |f(x+s) - f(x)|\varphi_y(s)\,ds \leq \int_{-\delta}^{\delta} \epsilon\varphi_y(s)\,ds \leq \int_{\mathbb{R}} \epsilon\varphi_y(s)\,ds = \epsilon.$$

For the second integral, note that for $|s| > \delta$,

$$\varphi_y(s) = \frac{1}{\pi}\frac{y}{s^2+y^2} \leq \frac{1}{\pi}\frac{y}{\delta^2+y^2} \leq \frac{y}{\pi\delta^2} \to 0 \text{ as } y \downarrow 0.$$

Hence, we can take the limit

$$\lim_{y\downarrow 0} \int_{|s|\geq\delta} |f(x+s) - f(x)|\varphi_y(s)\,ds = \int_{|s|\geq\delta} |f(x+s) - f(x)| \lim_{y\downarrow 0} \varphi_y(s)\,ds = 0.$$

(The interchange of the limit and the integral is justified by the Dominated Convergence Theorem, if you want to get technical.)

Combining all this with (2.26), we have thus shown that, for any $\epsilon > 0$,

$$\lim_{y\downarrow 0} \left| -\frac{1}{\pi}\operatorname{Im} G(x+iy) - f(x) \right| \leq \epsilon.$$

If this limit is $\leq \epsilon$ for any positive number $\epsilon > 0$, it cannot be anything other than 0, as we hoped to show. □

REMARK 2.41. The probability density φ_y, a rescaling of the Cauchy density, plays a key role above; this probability density is the "kernel" of the imaginary part of G. For that reason, some authors call G the *Cauchy transform* instead of the Stieltjes transform.

REMARK 2.42. What is really going on here is that the probability density φ_y is getting closer and closer to a "delta function" as $y \downarrow 0$: it is concentrating all of its mass at 0, and so in the limit, integrating against it just picks off the value of the function at 0. We've done the change of variables, so we're integrating $f(x+s)$ against $\varphi_y(s)$; this integral therefore approaches $f(x+0) = f(x)$.

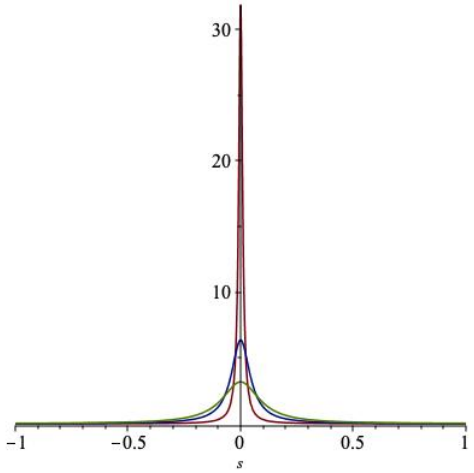


FIGURE 2.4. The graph of the rescaled Cauchy density $\varphi_y(s) = \frac{1}{\pi}\frac{y}{s^2+y^2}$ for $y = 0.1$, $y = 0.05$, and $y = 0.01$.

Great! We now have a robust tool, the Stieltjes transform, from which we can effectively recover the density. But this is only useful if we can, in fact, compute the Stieltjes transform! Remember our game: we have computed the moments of a so-far-unknown distribution. The question is: can we compute the Stieltjes transform $G(z)$ of this distribution only knowing its moments? The answer is, once again, *yes*, at least informally, by expanding a power series. Referring to Definition 2.37, if Y is a random variable with moments $\mathbb{E}(Y^k) = \mu_k$, then

$$G_Y(z) = \mathbb{E}\left(\frac{1}{z-Y}\right) = \frac{1}{z}\mathbb{E}\left(\frac{1}{1-Y/z}\right) = \frac{1}{z}\mathbb{E}\left(\sum_{k=0}^{\infty}(Y/z)^k\right).$$

Assuming we can exchange the sum and the expectation, this gives

$$G_Y(z) = \frac{1}{z}\sum_{k=0}^{\infty}\mathbb{E}((Y/z)^k) = \sum_{k=0}^{\infty}\frac{1}{z^{k+1}}\mathbb{E}(Y^k) = \sum_{k=0}^{\infty}\frac{\mu_k}{z^{k+1}}. \tag{2.27}$$

Thus: G_Y is a simple power-series involving the moments! That is, of course, if we can justify the above interchange of sum and $\mathbb{E}$. As with the moment generating function, the key is (uniform enough) convergence of the power series. From the root test, or the ratio test, we will therefore have the more rigorous statement that: if there is a constant $R > 0$ with $|\mu_k| \leq R^k$, then (2.27) holds true whenever $|z| > R$. That's not quite good enough for our purposes though: in order to use the Stieltjes inversion formula of Theorem 2.40, we need to know the value of $G_Y(z)$ for z arbitrarily close to each point in $\mathbb{R}$, not just those z outside some potentially large disk. In practice, the power series (2.27) will converge to a (complex analytic) function that is defined on all of $\mathbb{C}_+$, and there will be no problem. (Fully justifying this requires some complex analysis; we will content ourselves with concrete computations going forward.)

To summarize: we now have a tool to find the density corresponding to our moments: we will compute the power-series (2.27) for the moments μ_k in Theorem 2.32, and then apply the Stieltjes inversion formula of Theorem 2.40 to recover the density.

2.6. The Marčenko–Pastur Distribution

Following Corollary 2.23 and Theorem 2.32, we now have the tools to compute the asymptotic probability densities of the Wishart ensembles. The asymptotic moments are given in Theorem 2.32 as

$$\mu_k = \sum_{r=1}^{k} \frac{1}{r} \binom{k}{r-1} \binom{k-1}{r-1} \varrho^r. \tag{2.28}$$

They depend on the parameter ϱ, the asymptotic aspect ratio of the $m \times N$ feature matrix $\mathbf{X}$ for which the Wishart matrix is $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$; so we will label the asymptotic density as f_ϱ. To determine this density, we will first compute the Stieltjes transform G_ϱ associated to this unknown density:

$$G_\varrho(z) = \int_{\mathbb{R}} \frac{f_\varrho(t)}{z-t}\, dt = \sum_{k=0}^{\infty} \frac{\mu_k}{z^{k+1}}$$

cf. (2.27).

One approach would be to now substitute in the moments μ_k from (2.28), and try to sum this series directly. One might hope to interchange the order of the summations and use some nice binomial coefficient identities to come up with an explicit formula for the sum. This approach does not really work, unfortunately. Instead, we will use a different characterization of the moments μ_k which is actually an intermediate step on the way to finding the explicit formula (2.28) (which we did not discuss in Section 2.4, but would have come up from the combinatorial arguments that lead to the explicit formula for the moments.

PROPOSITION 2.43. *The moments μ_k of* (2.28) *are uniquely characterized by the following double recursion: $\mu_0 = \vartheta_0 = 1$ and, for $k \geq 1$,*

$$\mu_k = \varrho \sum_{j=1}^{k} \vartheta_{k-j}\mu_{j-1} \qquad \vartheta_k = \sum_{j=1}^{k} \mu_{k-j}\vartheta_{j-1}. \tag{2.29}$$

The auxiliary number ϑ_k are also determined by this recursion, but do not have a direct meaning to our original problem. These two numbers arise as weighted sums of the number of steps taken by certain walk on graphs that arise from the Wishart moment problem.

In order to use Proposition 2.43 to calculate G_ϱ, we're going to introduce two new "moment generating functions" associated to the two sequences $(\mu_k)_{k=0}^\infty$ and $(\vartheta_k)_{k=0}^\infty$.

$$M(\zeta) = \sum_{i=0}^{\infty} \mu_i \zeta^i \qquad T(\zeta) = \sum_{j=0}^{\infty} \vartheta_j \zeta^j. \tag{2.30}$$

These two power series will converge for $|\zeta|$ small, provided the sequences $(\mu_i)_{i=0}^\infty$ and $(\vartheta_j)_{j=0}^\infty$ don't grow faster than exponentially; this fact will come out of the analysis, so the following computations will be well-justified. Note that, if we can compute the function $M(\zeta)$ explicitly, then our task (to compute G_ϱ) is complete, because

$$G_\varrho(z) = \tfrac{1}{z} M(\tfrac{1}{z}). \tag{2.31}$$

Great. So, how do we compute $M(\zeta)$? The key is to notice that the recursive relationship between the coefficients μ_i and ϑ_j is familiar: it comes from power series multiplication. If we multiply

$M(\zeta)$ and $T(\zeta)$, we get

$$M(\zeta)T(\zeta) = \left(\sum_{i=0}^{\infty} \mu_i \zeta^i\right)\left(\sum_{j=0}^{\infty} \vartheta_j \zeta^j\right) = \sum_{i,j=0}^{\infty} \mu_i \vartheta_j \zeta^{i+j}.$$

This is still a power series in ζ, so it makes sense to label it by the powers of ζ; hence, we introduce a new index $n = i + j$:

$$M(\zeta)T(\zeta) = \sum_{n=0}^{\infty} \sum_{\substack{i,j\geq 0\\ i+j=n}} \mu_i \vartheta_j \zeta^{i+j} = \sum_{n=0}^{\infty} \zeta^n \sum_{\substack{i,j\geq 0\\ i+j=n}} \mu_i \vartheta_j.$$

The inside double sum is really just the sum over $i \in \{0, 1, \dots, n\}$, where $j = n - i$:

$$M(\zeta)T(\zeta) = \sum_{n=0}^{\infty} \zeta^n \sum_{i=0}^{n} \mu_i \vartheta_{n-i}. \tag{2.32}$$

The inside sum giving the coefficient of ζ^n is very close to the first sum in (2.29). Indeed, to compare:

$$\sum_{j=1}^{k} \vartheta_{k-j}\mu_{j-1} = \vartheta_{k-1}\mu_0 + \vartheta_{k-2}\mu_1 + \vartheta_{k-3}\mu_2 + \cdots + \vartheta_0\mu_{k-1}$$

while

$$\sum_{i=0}^{n} \mu_i \vartheta_{n-i} = \mu_0\vartheta_n + \mu_1\vartheta_{n-1} + \cdots + \mu_{n-1}\vartheta_1 + \mu_n\vartheta_0.$$

Matching these two, we see that they are identical if we set $n = k - 1$. Since the former is (up to a factor of ϱ) the coefficient of ζ^k in the power series $M(\zeta)$, we do the following:

$$M(\zeta)T(\zeta) = \frac{1}{\zeta}\sum_{n=0}^{\infty} \zeta^{n+1} \sum_{i=0}^{n} \mu_i \vartheta_{n-i} = \frac{1}{\zeta}\sum_{k=1}^{\infty} \zeta^k \sum_{i=0}^{k-1} \mu_i \vartheta_{k-1-i}$$

where we have reindexed $k = n + 1$. For the inside sum, if we reindex $k = i + 1$, we recover

$$\sum_{i=0}^{k-1} \mu_i \vartheta_{k-1-i} = \sum_{j=1}^{k} \mu_{j-1}\vartheta_{k-j}$$

as we'd hoped. The conclusion, now using (2.29), is that

$$M(\zeta)T(\zeta) = \frac{1}{\zeta}\sum_{k=1}^{\infty} \zeta^k \sum_{j=1}^{k} \mu_{j-1}\vartheta_{k-j} = \frac{1}{\zeta}\sum_{k=1}^{\infty} \zeta^k \cdot \frac{1}{\varrho}\mu_k.$$

This sum on the right-hand-size is almost the definition of $M(\zeta)$! We just need to throw in one missing term. Multiplying by $\varrho\zeta$ on both sides, we have

$$\varrho\zeta M(\zeta)T(\zeta) = \sum_{k=1}^{\infty} \mu_k \zeta^k = M(\zeta) - 1. \tag{2.33}$$

This gives us an equation for $M(\zeta)$, but involving the also-unknown $T(\zeta)$. However, we can do nearly the same analysis to find such an equation for $T(\zeta)$ (involving $M(\zeta)$). Going back to (2.32),

if we relabel the inside sum there with $j = n - i$, we have

$$M(\zeta)T(\zeta) = \sum_{n=0}^{\infty} \zeta^n \sum_{j=0}^{n} \mu_{n-j}\vartheta_j$$

and now the inside sum resembles the second sum in (2.29), which defined ϑ_k. Now following the above steps word-for-word, with the only difference being the lack of a factor of ϱ, we find that

$$\zeta M(\zeta)T(\zeta) = T(\zeta) - 1. \tag{2.34}$$

Combining the two equations (2.33) and (2.34), we can actually solve for $M(\zeta)$ (and $T(\zeta)$).

PROPOSITION 2.44. *The power series $M(\zeta)$ of* (2.30) *satisfies the quadratic equation*

$$\zeta M(\zeta)^2 + ((\varrho - 1)\zeta - 1)M(\zeta) + 1 = 0. \tag{2.35}$$

PROOF. Comparing (2.33) and (2.33), we see that $M(\zeta) - 1 = \varrho(T(\zeta) - 1)$, and therefore $\varrho T(\zeta) = M(\zeta) + \varrho - 1$. Substituting this into (2.33), we have

$$\zeta M(\zeta)(M(\zeta) + \varrho - 1) = M(\zeta) - 1$$

and this simplifies to (2.35). □

We could now solve this quadratic equation for $M(\zeta)$, but remembering that our goal is to find the function $G_\varrho(z)$, it will be neater to substitute back using (2.31) first. Letting $z = \frac{1}{\zeta}$ in (2.35), we have

$$\tfrac{1}{z}M(\tfrac{1}{z})^2 + ((\varrho - 1)\tfrac{1}{z} - 1)M(\tfrac{1}{z}) + 1 = 0.$$

Since $M(\frac{1}{z}) = zG_\varrho(z)$, this says

$$\frac{1}{z}(zG_\varrho(z))^2 + ((\varrho - 1)\frac{1}{z} - 1) \cdot zG_\varrho(z) + 1 = 0.$$

Simplifying this, we can now find an explicit formula for the desired Steiltjes transform $G_\varrho(z)$.

COROLLARY 2.45. *The Stieltjes transform G_ϱ satisfies the quadratic equation*

$$zG_\varrho(z)^2 + (\varrho - 1 - z)G_\varrho(z) + 1 = 0. \tag{2.36}$$

The solution can be expressed as follow. Let

$$\varrho_\pm = (1 \pm \sqrt{\varrho})^2. \tag{2.37}$$

Then

$$G_\varrho(z) = \frac{z - \varrho + 1 - \sqrt{(z - \varrho_-)(z - \varrho_+)}}{2z}. \tag{2.38}$$

PROOF. This is just the quadratic formula. The discriminant (under the root) is

$$(\varrho - 1 - z)^2 - 4z = z^2 - 2(1 + \varrho)z + (1 - \varrho)^2$$

and this factors as $(z - \varrho_-)(z - \varrho_+)$ as above. □

REMARK 2.46. We are accustomed to write $\pm\sqrt{}$ in the quadratic formula. In this case, it doesn't make so much sense to write $\pm$ since it is the square root of a complex number; while there are two roots, they are not $\pm$ some positive real number. We have chosen to use the minus sign here, with a bit of fore-knowledge that this will work out to give the correct answer in the next calculation.

We are *finally* in a position to calculate the exact (limit) density of eigenvalues for Wishart ensembles. We need only apply the Stieltjes inversion Theorem 2.40 to the function G_ϱ we've just finished calculating. The result is as follows.

THEOREM 2.47. *Let $\varrho \geq 1$, and let $(m_N)_{N=1}^\infty$ be a sequence of positive integers satisfying $\lim_{N\to\infty} \frac{m_N}{N} = \varrho$. Let $\mathbf{W} = \mathbf{W}^{m_N,N}$ be a Wishart matrix of size $m_N \times N$, with eigenvalue ensemble $(\boldsymbol{\lambda}^{(N)})$. Then for any $a < b$,*

$$\lim_{N\to\infty} h_{[a,b]}(\boldsymbol{\lambda}^{(N)}) = \lim_{N\to\infty} \frac{1}{N}\#\{j\colon \lambda_j^{(N)} \in [a,b]\} = \int_a^b f_\varrho(x)\,dx$$

where

$$f_\varrho(x) = \frac{\sqrt{(x-\varrho_-)(\varrho_+ - x)}}{2\pi x}\mathbb{1}_{[\varrho_-,\varrho_+]}(x) \tag{2.39}$$

where $\varrho_\pm$ are defined in (2.37)*. This function is called the* **Marčenko–Pastur** *density.*

PROOF. Following Corollaries 2.23, 2.33, and 2.45, we must simply apply the Stieltjes inversion formula to the function G_ϱ in (2.38):

$$f_\varrho(x) = -\frac{1}{\pi}\lim_{y\downarrow 0} \operatorname{Im} G_\varrho(x+iy).$$

To evaluate this limit, as usual, it is easiest to realize the denominator: multiplying by $\overline{z}$ in numerator and denominator of (2.38), and using $z = x + iy$, we have

$$\begin{aligned} G_\varrho(z) &= \frac{|z|^2 + (1-\varrho)\overline{z} - \overline{z}\sqrt{(z-\varrho_-)(z-\varrho_+)}}{2|z|^2} \\ &= \frac{1}{2} + \frac{(1-\varrho)(x-iy)}{2(x^2+y^2)} - \frac{x-iy}{2(x^2+y^2)}\sqrt{(z-\varrho_-)(z-\varrho_+)}. \end{aligned}$$

Now we must take $-\frac{1}{\pi}$ times the imaginary part. The first two terms are:

$$-\frac{1}{\pi}\operatorname{Im}\left(\frac{1}{2} + \frac{(1-\varrho)(x-iy)}{2(x^2+y^2)}\right) = (1-\varrho)\frac{y}{2\pi(x^2+y^2)}. \tag{2.40}$$

Here, provided $x \neq 0$, the limits as $y \downarrow 0$ is 0, so we can ignore this term (so long as there is no mass at $x = 0$). For the last term, we have a product of two fully complex numbers. Noting that $\operatorname{Im}[(a+ib)(c+id)] = ad + bc$, we have

$$\begin{aligned} &-\frac{1}{\pi}\operatorname{Im}\left(-\frac{x-iy}{2(x^2+y^2)}\sqrt{(z-\varrho_-)(z-\varrho_+)}\right) \\ =&\frac{1}{2\pi(x^2+y^2)}\operatorname{Im}\left((x-iy)\sqrt{(z-\varrho_-)(z-\varrho_+)}\right) \\ =&\frac{1}{2\pi(x^2+y^2)}\left(x\cdot\operatorname{Im}\left(\sqrt{(z-\varrho_-)(z-\varrho_+)}\right) - y\cdot\operatorname{Re}\left(\sqrt{(z-\varrho_-)(z-\varrho_+)}\right)\right). \end{aligned} \tag{2.41}$$

We can now take $y \downarrow 0$. Again, provided $x \neq 0$, $\frac{y}{2\pi(x^2+y^2)} \to 0$, and so the second term vanishes in the limit. For the first term, as $y \downarrow 0$, $\sqrt{(z-\varrho_-)(z-\varrho_+)} \to \sqrt{(x-\varrho_-)(x-\varrho_+)}$. The quadratic under the square root has zeroes at $x = \varrho_\pm$, and is a convex quadratic; hence, it is negative for $x \in (\varrho_-, \varrho_+)$ and positive for x outside this interval. Thus means that

$$\operatorname{Im}\left(\sqrt{(x-\varrho_-)(x-\varrho_+)}\right) = \sqrt{-(x-\varrho_-)(x-\varrho_+)}\mathbb{1}_{[\varrho_-,\varrho_+]}(x).$$

Sine $\frac{x}{2\pi(x^2+y^2)} \to \frac{1}{2\pi x}$, this yields (2.39), and concludes the proof. □

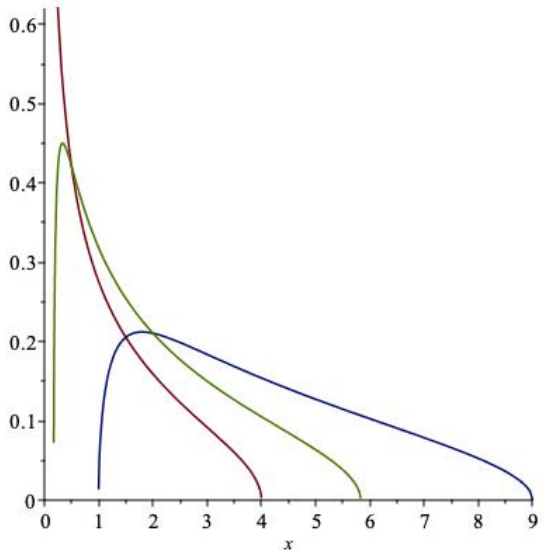


FIGURE 2.5. The graphs of the Marčenko–Pastur density f_ϱ for $\varrho = 1, 2, 4$.

As seen in Figure 2.5, f_1 has a vertical asymptote at $x = 0$; for $\varrho > 0$ the lower bound of the support $\varrho_- = (1 - \sqrt{\varrho}^2)$ is strictly positive and the denisty is continuous everywhere. There are some numerical artifacts in the figure, resulting in the density not being fully drawn near ϱ_- in each case; the function $f_\varrho(x)$ does come smoothly and sharply down to 0 as $x \downarrow \varrho_+$.

We have now identified the "shape of random noise": if the data points $\mathbf{x}_j$ (the columns of $\mathbf{X}$) are all (centered) independent random vectors, with independent entries, then the histogram of nonzero eigenvalues of the sample covariance matrix $\frac{1}{N}\mathbf{X}\mathbf{X}^\top$ (which are the same as the nonzero eigenvalues of $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$) has a deterministic shape that depends only on the aspect ratio ϱ of the feature matrix $\mathbf{X}$. In particular, the top end of the histogram of eigenvalues ends at $\varrho_+ = (1 + \sqrt{\varrho})^2$.

This suggests that, in a real-world data set (of dimension m greater than the number N of sample points), when doing PCA, we should look only for eigenvalues of the sample covariance matrix that are bigger than ϱ_+. We should be careful, however: the setup above does not necessarily mean that a pure noise matrix cannot have eigenvalues larger than this, even as $N \to \infty$. After all, what we have proved is the following: if $a > \varrho_+$, then

$$\lim_{N\to\infty} \frac{1}{N} \#\{j \colon \lambda_j^{(N)} > a\} = \int_a^\infty f_\varrho(x)\, dx = 0. \tag{2.42}$$

So the *proportion* of eigenvalues that are larger than ϱ_+ is asymptotically 0. But that doesn't prevent the *largest* eigenvalue, or even the largest *million* eigenvalues from being larger in the limit. Indeed, let $L^{(N)} = \#\{j \colon \lambda_j^{(N)} > \varrho_+\}$; then (2.42) tells us that

$$\lim_{N\to\infty} \frac{L^{(N)}}{N} = 0.$$

If $L^{(N)} = \max\{N, 10^6\}$, or even $L^{(N)} = \sqrt{N}$, this is perfectly consistent with what we've proved.

In practice, if you start doing simulations of the largest eigenvalue $\lambda_1^{(N)}$ of a Wishart ensemble, you will see the following stronger fact holds true.

THEOREM 2.48 (Bai, Yin, 1993). *Under the same conditions as in Theorem 2.47, with the additional assumption that the entries $[\mathbf{X}]_{ij}$ of the (random noise) feature matrix have finite 4th moments, we have*

$$\lim_{N\to\infty} \lambda_1^{(N)} = \varrho_+ \qquad \textit{and} \qquad \lim_{N\to\infty} \lambda_N^{(N)} = \varrho_-$$

with probability 1. *That is: the largest and smallest eigenvalues converge almost surely to the edges of the support of the Marčenko–Pastur density.*

This theorem was a major achievement, published in the Annals of Probability. The method of proof is beyond the tools we have developed, although it is in the same ball-park. We saw in Section 2.4 that, to understand the *bulk density* of eigenvalues, it sufficed to compute the large-N limits of $\mathrm{tr}_N(\mathbf{W}^k)$. A similar but more involved argument, invented by Bai and Yin, shows that to understand the *extreme* (largest and smallest) eigenvalues, it suffices to understand the large-N limits of $\mathrm{tr}_N(\mathbf{W}^{k_N})$, where now the power $k = k_N$ must be allowed to grow as N grows (at least logarithmically fast) toward ∞. The resulting proof thus delves even deeper into combinatorial structures that arise from these expansions. We will not discuss this further at present, and will satisfy ourselves with the truth of Theorem 2.48.

Let's investigate the density f_ϱ further, particularly in the case that $\varrho < 1$. After all, we have been making the assumption that $m \geq N$, but our analysis above didn't seem to explicitly rely on this assumption. As the next computation shows, something strange does happen as ϱ falls below 1. It is fairly evident that f_ϱ is a non-negative function for any $\varrho > 0$; however, we have the following.

PROPOSITION 2.49. *Let f_ϱ be the function defined in* (2.39)*, for any $\varrho > 0$. Then*

$$\int_{\mathbb{R}} f_\varrho(x)\,dx = \int_{\varrho_-}^{\varrho_+} f_\varrho(x)\,dx = \begin{cases} 1, & \varrho \geq 1 \\ \varrho, & \varrho \in (0,1). \end{cases}$$

PROOF. It turns out that the antiderivative of f_ϱ can be explicitly computed, with some clever tricks. For any positive $a < b$, let us introduce the function $r(x) = (x-a)(b-x)$. Then $f_\varrho(x)$ is a constant multiple of $\sqrt{r(x)}/x$ with $a = \varrho_-$ and $b = \varrho_+$. Now, notice the following:

$$\begin{aligned} r(x) &= -x^2 + (a+b)x - ab \\ &= x(-x + (a+b)) - ab \end{aligned}$$

while $r'(x) = -2x + (a+b)$. By comparison,

$$\frac{2}{x}(r(x) + ab) = \frac{2}{x} \cdot x(-x + (a+b)) = -2x + 2(a+b).$$

Thus

$$r'(x) = \frac{2}{x}(r(x) + ab) - (a+b)$$

which we can rearrange to give

$$\frac{r'(x)}{2} = \frac{r(x)}{x} + \frac{ab}{x} - \frac{a+b}{2}.$$

Now restricting to $x \in (a,b)$ where $r(x) > 0$, we can divide through by $\sqrt{r(x)}$ to yield

$$\frac{r'(x)}{2\sqrt{r(x)}} = \frac{\sqrt{r(x)}}{x} + \frac{ab}{x\sqrt{r(x)}} - \frac{a+b}{2\sqrt{r(x)}}.$$

The left-hand-side is the derivative of $\sqrt{r(x)}$, and so we find that

$$\sqrt{r(x)} = \int \frac{\sqrt{r(x)}}{x}\,dx + ab\int \frac{dx}{x\sqrt{r(x)}} - \frac{a+b}{2}\int \frac{dx}{\sqrt{r(x)}}$$

or, rearranging,

$$\int \frac{\sqrt{r(x)}}{x}\,dx = \sqrt{r(x)} + \frac{a+b}{2}\int \frac{dx}{\sqrt{r(x)}} - ab\int \frac{dx}{x\sqrt{r(x)}}. \tag{2.43}$$

The two integrals on the right-hand-side of (2.43) can each be evaluated using standard trigonometric substitutions.

$$\int \frac{dx}{\sqrt{r(x)}} = \arcsin\left(\frac{2x-(a+b)}{b-a}\right) \tag{2.44}$$

$$\int \frac{dx}{x\sqrt{r(x)}} = \frac{1}{\sqrt{ab}}\arcsin\left(\frac{(a+b)x-2ab}{x(b-a)}\right). \tag{2.45}$$

Combining (2.43) with (2.44) and (2.45) gives an explicit formula for the antiderivative of $\sqrt{r(x)}/x$. We only care about the integral over the interval $[a,b]$, so we can just evaluate. Note that $r(a) = r(b) = 0$, so the first term in (2.43) doesn't contribute; we therefore have

$$\begin{aligned}\int_a^b \frac{\sqrt{r(x)}}{x}\,dx = \frac{a+b}{2}&\left[\arcsin\left(\frac{2b-(a+b)}{b-a}\right) - \arcsin\left(\frac{2a-(a+b)}{b-a}\right)\right]\\ &- \sqrt{ab}\left[\arcsin\left(\frac{(a+b)b-2ab}{b(b-a)}\right) - \arcsin\left(\frac{(a+b)a-2ab}{a(b-a)}\right)\right].\end{aligned}$$

The terms inside the $\arcsin$'s are ± 1, and so the differences becomes $\frac{\pi}{2} - (-\frac{\pi}{2}) = \pi$; thus

$$\int_a^b \sqrt{r(x)}x\,dx = \pi\left(\frac{a+b}{2} - \sqrt{ab}\right).$$

Finally, specializing the the Marčenko–Pastur density: we have $a = \varrho_-$, $b = \varrho_+$, and $f_\varrho(x) = \frac{\sqrt{r(x)}}{2\pi x}$; therefore

$$\int_{\varrho_-}^{\varrho_+} f_\varrho(x)\,dx = \frac{1}{2\pi}\cdot\pi(\varrho_- + \varrho_+ - \sqrt{\varrho_-\varrho_+}).$$

Since $\varrho_\pm = (1 \pm \sqrt{\varrho})^2$,

$$\varrho_- + \varrho_+ = (1 + 2\sqrt{\varrho} + \varrho) + (1 - 2\sqrt{\varrho} + \varrho) = 2(1+\varrho)$$

and

$$\varrho_-\varrho_+ = [(1-\sqrt{\varrho})(1+\sqrt{\varrho})]^2 = (1-\varrho)^2. \tag{2.46}$$

Therefore $\sqrt{\rho_-\rho_+} = |1-\varrho|$, and so

$$\int_{\varrho_-}^{\varrho_+} f_\varrho(x)\,dx = \frac{1}{2}(1 + \varrho - |1-\varrho|).$$

This establishes the result. □

Thus, when $\varrho < 1$, f_ϱ is *not a probability density*! This seems very problematic: how can it describe the asymptotic density if eigenvalues if it is not a density? To understand what happened, let's return to the matrix $\mathbf{W} = \mathbf{W}^{m,N}$. When $\varrho < 1$, this means that $\frac{m}{N} < 1$ (for large m and N). Now, $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$, and $\mathbf{X}$ is $m \times N$. Thus $\mathbf{W}$ is $N \times N$; but since $m \leq N$, the rank of $\mathbf{X}$, and therefore the rank of $\mathbf{W}$, cannot exceed m. This means that all but at most m eigenvalues of $\mathbf{W}$ must be 0. In other words:

When $m \leq N$, the proportion of eigenvalues of $\mathbf{W}$ that are equal to 0 is at least $1 - \frac{m}{N}$.

Since we are dealing with the scaling regime where $\frac{m}{N} \to \varrho$, this means that we expect to see a *spike* in the histogram: there is a point mass of weight at least $1 - \varrho$ at 0. In other words: there is no *density* of eigenvalues; the distribution of eigenvalues is a probability distribution that is not continuous. But it isn't far off: it mostly has density f_ϱ, but it also has a spike at 0; so we might (somewhat informally) write the "density" of eigenvalues as

$$(1 - \varrho)\delta_0 + f_\varrho. \tag{2.47}$$

Here, δ_0 stands for a unit mass at 0; in other words, it is the (discrete) distribution of a "random" variable that takes value 0 with probability 1. So (2.47) is short hand for: *the random variable has probability $(1 - \varrho)$ of equaling 0, and is otherwise given by density ϱ* when $\varrho \in (0, 1)$. More pedantically, if we let F_ϱ denote the *cumulative distribution function*, the statement is that

$$F_\varrho(t) = \begin{cases} 0, & t < 0 \\ 1 - \varrho, & 0 \leq t < \varrho_- \\ 1 - \varrho + \int_{\varrho_-}^t f_\varrho(x)\,dx, & t \geq \varrho_- \end{cases} \qquad \text{for } \varrho < 1. \tag{2.48}$$

By comparison,

$$F_\varrho(t) = \begin{cases} 0, & t < \varrho_- \\ \int_{\varrho_-}^t f_\varrho(x)\,dx, & t \geq \varrho_- \end{cases} \qquad \text{for } \varrho \geq 1. \tag{2.49}$$

The question is: why did we not see this come out of our analysis for the Stieltjes transform, in Theorem 2.47? The answer is that we *did* see it, we were just a tiny bit careless. If you review the proof again, in particular (2.40), you see that all of the limit statements required the assumption $x \neq 0$ to really work. If we knew that the result would be a probability density, that didn't matter, since a probability density assigns 0 mass to any point. We now know that it was incorrect to assume there would be a density, at least in the case $\varrho < 1$. Looking again very carefully at (2.40) and (2.41), in the case $x = 0$, we see that for $y > 0$,

$$-\frac{1}{\pi}\mathrm{Im}\, G_\varrho(0 + iy) = (1 - \varrho)\frac{y}{2\pi(0^2 + y^2)} + \frac{-y \cdot \mathrm{Re}\left(\sqrt{(0 + iy - \varrho_-)(0 + iy - \varrho_+)}\right)}{2\pi(0^2 + y^2)}.$$

Now, note that

$$\lim_{y \downarrow 0} \sqrt{(0 + iy - \varrho_-)(0 + iy - \varrho_+)} = \sqrt{\varrho_-\varrho_+} = |1 - \varrho|$$

from (2.46). Therefore (being a little fast and loose with limits here), we should expect that

$$-\frac{1}{\pi}\lim_{y \downarrow 0}\mathrm{Im}\, G_\varrho(0 + iy) = \frac{1}{2\pi}\left(1 - \varrho - |1 - \varrho|\right)\lim_{y \downarrow 0}\frac{y}{0 + y^2}.$$

The coefficient should have come out to $1 - \varrho + |1 - \varrho|$; there is a sign error somewhere and I haven't been able to find it. When $\varrho \geq 1$, $1 - \varrho + |1 - \varrho| = 0$, and so the analysis in the proof of Theorem 2.47 is completely correct (confirmed by the fact that f_ϱ really is a probability density

in that case). But when $\varrho < 1$, $\frac{1}{2\pi}(1 - \varrho - |1 - \varrho|) = \frac{1}{\pi}(1 - \varrho)$. This is where the additional point-mass at $x = 0$ comes from in this case, from the Stieltjes transform analysis.

REMARK 2.50. Of course, $\lim_{y\downarrow 0} \frac{y}{0+y^2} = \lim_{y\downarrow 0} \frac{1}{y}$ does not actually exist; it diverges to $+\infty$. It is not literally true that $-\frac{1}{\pi}G_\varrho(x+iy)$ converges, as a function, to $(1-\varrho)\delta_0(x) + f_\varrho(x)$ as $y \downarrow 0$. If we were being very precise, we would be using the notion of "weak convergence", which identifies limits not directly but when integrated against "test functions". If one performs that analysis, one doesn't literally have $\frac{1}{\pi}\frac{y}{0+y^2}$ here but rather $\frac{1}{\pi}\frac{y}{s^2+y^2}$ for a new auxiliary integration variable s. Recall that this is what we called the *Cauchy density* $\varphi_y(s)$ in the proof of Theorem 2.40, and in that proof we showed that $\int \varphi_y(s)\psi(s)\,ds \to \psi(0)$ as $y \downarrow 0$. That is the proper sense of convergence to δ_0.

Let's conclude this section, and chapter, but fully stating the Marčenko–Pastur result, both for Wishart matrices $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$ and the corresponding sample covariance matrices $\mathbf{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top$.

THEOREM 2.51. *Let $\varrho > 0$, and let f_ϱ be the function defined in* (2.39)*. Let $(m_N)_{N=1}^\infty$ be a sequence of positive integers satisfying $\lim_{N\to\infty} \frac{m_N}{N} = \varrho$. Let $\mathbf{X}_N$ be an $m_N \times N$ feature matrix whose entries are all i.i.d. $\mathscr{N}(0,1)$ random variables. Define*

$$\mathbf{W}_N = \frac{1}{N}\mathbf{X}_N^\top\mathbf{X}_N \qquad \textit{and} \qquad \mathbf{C}_N = \frac{1}{N}\mathbf{X}_N\mathbf{X}_N^\top.$$

Let $a < b$. As $N \to \infty$, the proportion of eigenvalues of $\mathbf{W}_N$ in $[a,b]$ converges to

$$\begin{cases} \displaystyle\int_a^b f_\varrho(x)\,dx, & \text{if } \varrho \geq 1 \\ \displaystyle(1-\varrho)\mathbb{1}_{[a,b]}(0) + \int_a^b f_\varrho(x)\,dx, & \text{if } \varrho \in (0,1). \end{cases}$$

Also, as $N \to \infty$, the proportion of eigenvalues of $\mathbf{C}_N$ in $[a,b]$ converges to

$$\begin{cases} \displaystyle\int_a^b \frac{1}{\varrho} f_\varrho(x)\,dx, & \text{if } \varrho \in (0,1) \\ \displaystyle(1-\tfrac{1}{\varrho})\mathbb{1}_{[a,b]}(0) + \int_a^b \tfrac{1}{\varrho} f_\varrho(x)\,dx, & \text{if } \varrho \geq 1. \end{cases}$$

PROOF. The statement for $\mathbf{W}_N$ was fully justified in the discussion on the last two pages. (The proportion in question can be written in terms of the cumulative distribution function F_ϱ as $F_\varrho(b) - F_\varrho(a)$. When the CDF has a jump, as this one does when $\varrho < 1$, this gives a term of the jump height when the jump point is in the interval, which we write as $(1-\varrho)\mathbb{1}_{[a,b]}(0)$.) For $\mathbf{C}_N$, Corollary 2.33 shows that the moments $\operatorname{tr}_N(\mathbf{W}_N^k)$ and $\operatorname{tr}_N(\mathbf{C}_N^k)$ are off by a constant factor of ϱ (when $k \geq 1$), which is what leads to the density being $\frac{1}{\varrho}f_\varrho$ in this case; and the structural point-mass at $x = 0$ is now in the case $\varrho > 1$ instead of $\varrho < 1$, due to the matrix $\mathbf{C}_N$ being $m \times m$ rather than $N \times N$. The details are left to the reader. □

CHAPTER 3

The BBP Phase Transition

In the last chapter, we dealt with the baseline situation that our $m \times N$ feature matrix $\mathbf{X}$ is completely random: all the columns $\mathbf{x}_j \in \mathbb{R}^m$ are independent from each other (representing the usual model of random sampling), and each of the N random vectors $\mathbf{x}_j$ is distributed as a standard normal $\mathcal{N}(\mathbf{0}, I_m)$. We found a universal profile of the eigenvalues of the sample covariance matrix $\frac{1}{N}\mathbf{X}\mathbf{X}^\top$ in this case, the Marčenko–Pastur distribution.

We now wish to consider a model of a more realistic situation: where there actually is some structure/signal in the data. We have already considered the model (Problem 4 on the Midterm) where $\mathbf{x}_j = \mathbf{v}_j + \mathbf{Z}_j$ where $\mathbf{v}_j$ are fixed (deterministic) vectors and $\mathbf{Z}_j$ are independent Gaussian noise (and we saw this model led to PCA as the MLE for estimating the subspace spanned by the $\mathbf{v}_j$). However, this model doesn't quite fit many data sets, where we expect that the data points are all independent samples of the *same* distribution. If we wish to enforce that assumption with the above model, this would require all the deterministic points $\mathbf{v}_j$ to be equal to some fixed vector $\mathbf{v}$. The problem with this is that then the sample mean would be $\overline{\mathbf{x}}_N = \mathbf{v} + \overline{\mathbf{Z}}_N$, and so we would have

$$\mathbf{x}_j^\circ = \mathbf{x}_j - \overline{\mathbf{x}}_N = (\mathbf{v}_j + \mathbf{Z}_j) - (\mathbf{v} + \overline{\mathbf{Z}}_N) = \mathbf{Z}_j - \overline{\mathbf{Z}}_N = \mathbf{Z}_j^\circ.$$

That is: after centering, there is no dependence on $\mathbf{v}$, and so the sample covariance matrix is the same as the pure noise case: the eigenvalues will follow the Marčenko–Pastur distribution, with the largest and smallest eigenvalues sticking to the edge as in Theorem 2.48.

3.1. Spiked Covariance Models

How can we model structure/signal in independent samples? The idea is to allow the signal to be a little noisy as well. After all, there are probably several sources of noise: environmental, measurement error, and perhaps some inherent randomness in the underlying signal itself. Therefore, we more accurately model the data as

$$\mathbf{x}_j = \mathbf{G}_j + \mathbf{Z}_j$$

where the $\mathbf{Z}_j$ are i.i.d. standard normal random vectors $\mathcal{N}(\mathbf{0}, I_m)$ (the environmental noise), and $\mathbf{G}_j$ is another Gaussian random vector, independent from $\mathbf{Z}_j$, this time with a possibly non-trivial mean $\boldsymbol{\mu}$ and covariance C:

$$\mathbf{G}_j \sim \mathcal{N}(\boldsymbol{\mu}, C), \qquad 1 \le j \le N.$$

We can immediately dispense with $\boldsymbol{\mu}$: as in the simpler model in the last paragraph, it will wash out after centering to form the sample covariance matrix. So we model $\mathbf{G}_j \sim \mathcal{N}(\mathbf{0}, C)$ for some non-identity covariance C. This is an $m \times m$ matrix; we are expecting the data to "live in" a lower-dimensional affine subspace, and this will be reflected in C being a low-rank matrix. (I.e. the signal / structure is encoded in C.)

To begin analyzing this model, we should note what is the true covariance of the random vector $\mathbf{x}_1$ (which is the same as for all the other $\mathbf{x}_j$). Denoting the components of the vector $\mathbf{x}_1$ as

$[x_1^1, \ldots, x_1^m]^\top$ (and similarly with G_1^i and Z_1^i), we have

$$\mathrm{Cov}(x_1^i, x_1^j) = \mathrm{Cov}(G_1^i + Z_1^i, G_1^j + Z_1^j) = \mathrm{Cov}(G_1^i, G_1^j) + \mathrm{Cov}(Z_1^i, Z_1^j)$$

where the cross-terms are 0 because $\mathbf{G}_1$ and $\mathbf{Z}_1$ are independent. By definition, this means

$$\mathrm{Cov}(\mathbf{x}_1) = \mathrm{Cov}(\mathbf{G}_1) + \mathrm{Cov}(\mathbf{Z}_1) = C + I_m.$$

Hence, our model is that the (true) covariance of each data point is a perturbation of the identity: $I + C$ where (hopefully) C is a low rank matrix. Let's formalize this.

DEFINITION 3.1. *Fix a positive integer r. A* **spiked covariance model** *of rank r is a random matrix $\mathbf{X} \in \mathbb{M}_{m\times N}$, where the columns of $\mathbf{X}$ are i.i.d. Gaussian random vectors with covariance $I_m + C_r$, where C_r is a positive semi-definite matrix of rank r.*

We would now like to study the sample covariance matrix (or corresponding Wishart matrix) of the given feature matrix $\mathbf{X}_N$. Before doing so, we note that a simplification comes quickly from Definition 3.1. The matrix C_r is symmetric, and so it can be diagonalized: $C_r = QD_rQ^\top$ where the eigenvector matrix Q is in $\mathrm{O}(m)$, and D_r has the form

$$D_r = \begin{bmatrix} \gamma_1 & & & & \\ & \gamma_2 & & & \\ & & \ddots & & \\ & & & \gamma_r & \\ & & & & \mathbf{0}_{m-r} \end{bmatrix} \tag{3.1}$$

with non-zero eigenvalues $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_r > 0$. The matrix $I_m + C_r$ and the matrix $I_m + D_r$ are similar. Indeed:

$$I_m + C_r = I_m + QD_rQ^\top = Q(I_m + D_r)Q^\top.$$

This observation basically allows us to discard C_r and replace it with D_r.

LEMMA 3.2. *If $\mathbf{X}$ is a spiked covariance model, with covariance $I_m + C_r$, and (Q, D_r) are defined as above, let $A = Q\Lambda Q^\top$ where Λ is the diagonal matrix with $[\Lambda]_{jj} = (1+\gamma_j)^{-1/2}$ for $j \leq r$ and $[\Lambda]_{jj} = 1$ for $j \geq r$. Then $\mathbf{Z} = A\mathbf{X}$ has i.i.d. $\mathcal{N}(\mathbf{0}, I_m)$ columns.*

The matrix A has been designed so that $A^2(I_m + C_r) = I_m$; i.e. $A = (I_m + C_r)^{-1/2}$. That is what is needed to make this lemma work.

PROOF. The jth column $\mathbf{Z}_j$ of $\mathbf{Z}$ is $A\mathbf{X}_j$; since the $\mathbf{X}_j$ are i.i.d., so are the $\mathbf{Z}_j$. It is a standard multivariate calculus exercise that any linear transformation of a (centered) Gaussian random vector is a (centered) Gaussian random vector, so $\mathbf{Z}_j$ is Gaussian with mean $\mathbf{0}$. It just behooves us to calculate its covariance matrix. This is a standard calculation. It suffices to work with $\mathbf{Z}_1$. For any $i, j \in [m]$,

$$\mathrm{Cov}([\mathbf{Z}_1]_i, [\mathbf{Z}_1]_j) = \mathrm{Cov}([A\mathbf{X}_1]_i, [A\mathbf{X}_1]_j).$$

By definition

$$[A\mathbf{X}_1]_i = \sum_{a=1}^m [A]_{ia}[\mathbf{X}_1]_a.$$

Therefore

$$\mathrm{Cov}([\mathbf{Z}_1]_i, [\mathbf{Z}_1]_j) = \mathrm{Cov}\left(\sum_{a=1}^m [A]_{ia}[\mathbf{X}_1]_a, \sum_{b=1}^m [A]_{jb}[\mathbf{X}_1]_b\right).$$

Since Cov is a bilinear form (i.e. linear in each of the two variable separately),

$$\mathrm{Cov}([\mathbf{Z}_1]_i, [\mathbf{Z}_1]_j) = \sum_{a,b=1}^{m} [A]_{ia}[A]_{jb}\mathrm{Cov}([\mathbf{X}_1]_a, [\mathbf{X}_1]_b).$$

The covariance matrix of $\mathbf{X}_1$ is $I + C_r$, and so we can write this as

$$\mathrm{Cov}([\mathbf{Z}_1]_i, [\mathbf{Z}_1]_j) = \sum_{a,b=1}^{m} [A]_{ia}[A]_{jb}[I_m + C_r]_{ab} = \sum_{a,b=1}^{m} [A]_{ia}[I_m + C_r]_{ab}[A^\top]_{bj} = [A(I_m + C_r)A^\top]_{ij}.$$

Now, employing the spectral decomposition of $A = A^\top$ and $I_m + C_r$, we have

$$A(I_m + C_r)A^\top = (Q\Lambda Q^\top)(Q(I_m + D_r)Q^\top)(Q\Lambda Q^\top) = Q\Lambda(I_m + D_r)\Lambda Q^\top.$$

Diagonal matrices commute, so this is equal to $Q\Lambda^2(I_m + D_r)Q^\top$. The matrix Λ was designed so that $\Lambda^2(I_m + D_r) = I_m$, and so the whole matrix is just equal to $QQ^\top = I_m$. Thus, the covariance matrix of $\mathbf{Z}_1$ is I_m; so $\mathbf{Z}_1 \sim \mathbf{Z}_j \sim \mathcal{N}(\mathbf{0}, I_m)$ as claimed. □

The upshot here is that, since $\mathbf{Z} = A\mathbf{X}$, we can express $\mathbf{X} = A^{-1}\mathbf{Z}$; the matrix A in the lemma is evidently invertible with $A^{-1} = Q\Lambda^{-1}Q^\top$ and $[\Lambda^{-1}]_{jj} = (1 + \lambda_j)^{1/2}$ for $j \leq r$ and $[\Lambda^{-1}]_{jj} = 1$ for $j > r$. That is, any spiked covariance model can be expressed in the form $A^{-1}\mathbf{Z}$, where $\mathbf{Z}$ has i.i.d. $\mathcal{N}(\mathbf{0}, I_m)$ columns: in other words, it is just a linear transformation composed with the underlying feature matrix of the Wishart ensembles of the last chapter.

Now, let's consider the corresponding Wishart matrices: $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$. (Again, we ignore the subtraction of the sample mean from the columns, since we assume N is large and since all the vectors are centered normal, by the Laws of Large Numbers the sample mean will be close to $\mathbf{0}$.) We simply observe that

$$\mathbf{X}^\top\mathbf{X} = (A^{-1}\mathbf{Z})^\top(A^{-1}\mathbf{Z}) = \mathbf{Z}^\top(A^{-1})^\top A^{-1}\mathbf{Z}.$$

Since A is symmetric, so is A^{-1}, and so

$$\mathbf{X}^\top\mathbf{X} = \mathbf{Z}^\top A^{-2}\mathbf{Z}.$$

But, by design, $A^{-2} = I_m + C_r$; and so we have

$$\mathbf{X}^\top\mathbf{X} = \mathbf{Z}^\top(I_m + C_r)\mathbf{Z}. \tag{3.2}$$

This is an easy way to construct any spiked covariance model. Moreover, together with the spectral decomposition, it gives us the following.

COROLLARY 3.3. *Let $\mathbf{X}$ be a spiked covariance model, with covariance $I_m + C_r$, and let $C_r = QD_rQ^\top$, as in* (3.1)*. Let $\mathbf{X}'$ be a spiked covariance model, with covariance $I_m + D_r$. Then the random matrices $\mathbf{X}^\top\mathbf{X}$ and $(\mathbf{X}')^\top\mathbf{X}'$ have the same distribution.*

PROOF. We simply continue the computation from above:

$$\mathbf{X}^\top\mathbf{X} = \mathbf{Z}^\top(I_m + C_r)\mathbf{Z} = \mathbf{Z}^\top Q(I_m + D_r)Q^\top\mathbf{Z} = (Q^\top\mathbf{Z})^\top(I_m + D_r)(Q^\top\mathbf{Z}).$$

Hence, to complete the proof, it suffices to show that the matrix $Q^\top\mathbf{Z}$ has the same distribution as $\mathbf{Z}$; i.e. it is an $m \times N$ matrix whose columns are i.i.d. $\mathcal{N}(\mathbf{0}, I_m)$ random vectors. To that end, note that the columns of $Q^\top\mathbf{Z}$ are just $Q^\top\mathbf{Z}_j$ (i.e. $Q^\top$ times the columns of $\mathbf{Z}$); since the columns of $\mathbf{Z}$ are i.i.d. therefore so are the columns of $Q^\top\mathbf{Z}$. Thus, it suffices to work with just the first column $\mathbf{Z}_1$: we must show that if $\mathbf{Z}_1$ is a standard Gaussian random vector in $\mathbb{R}^m$ and $Q \in \mathrm{O}(m)$, then $Q^\top\mathbf{Z}_1$ is also a standard random vector. For this we may appeal to the proof of Lemma 3.2: $Q^\top\mathbf{Z}_1$

is another centered Gaussian random vector, whose covariance matrix is $Q^\top (Q^\top)^\top = Q^\top Q = I_m$, concluding the proof. □

Thus, when considering spiked covariance models, we need never consider anything more general than a *diagonal* covariance: from the perspective of the corresponding Wishart matrix $\frac{1}{N}\mathbf{X}^\top\mathbf{X}$, the spiked covariance $I_m + C_r$ and the diagonal spiked covariance $I_m + D_r$ yield the same distribution, and hence will have eigenvectors and eigenvalues of the same distributions. Therefore, from this point forward, we restrict our attention to spiked covariance models with diagonal covariance perturbation of the form (3.1).

REMARK 3.4. This seems quite remarkable. After all, having diagonal covariance means that the data points, the columns of $\mathbf{X}$, still have all independent entries. So we have not really introduced any correlations at all – just rescaled the variances of the first r entries. How can this yield all the information in our signal? The answer is: it doesn't, but that's because we're looking at the Wishart ensemble $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$. Our original goal, performing PCA, requires finding the eigen*vectors* of the sample covariance matrix $\mathbf{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top$. While $\mathbf{W}$ and $\mathbf{C}$ have the same nonzero eigenvalues, their eigen*vectors* are quite different. The discussion above shows that the (random) eigenvectors of $\mathbf{W}' = \frac{1}{N}(\mathbf{X}')^\top\mathbf{X}'$ are the same, in distribution, as those of $\mathbf{W} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$. But the two matrices $\mathbf{C}' = \frac{1}{N}\mathbf{X}'(\mathbf{X}')^\top$ and $\mathbf{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top$ can have *very different* eigenvectors. Indeed, recalling that $\mathbf{X} = A^{-1}\mathbf{Z}$ with A symmetric,

$$N\mathbf{C} = \mathbf{X}\mathbf{X}^\top = A^{-1}\mathbf{Z}(A^{-1}\mathbf{Z})^\top = A^{-1}\mathbf{Z}\mathbf{Z}^\top A^{-1} = Q\Lambda^{-1}Q^\top\mathbf{Z}\mathbf{Z}^\top Q\Lambda^{-1}Q^\top. \tag{3.3}$$

On the other hand, $\mathbf{X}'$ can be constructed in the form $A^{-1}\mathbf{Z}$, where $A^2(I_m + D_r) = I_m$; this is precisely the Λ above. Hence

$$N\mathbf{C}' = \mathbf{X}'(\mathbf{X}')^\top = \Lambda^{-1}\mathbf{Z}\mathbf{Z}^\top\Lambda^{-1}.$$

As noted above, $\mathbf{Z} \sim Q^\top\mathbf{Z}$, and so

$$N\mathbf{C}' \sim \Lambda^{-1}(Q^\top\mathbf{Z})(Q^\top\mathbf{Z})^\top\Lambda^{-1} = \Lambda^{-1}Q^\top\mathbf{Z}\mathbf{Z}^\top Q^\top\Lambda^{-1}.$$

Comparing this to (3.3), we see that $\mathbf{C}$ and $\mathbf{C}'$ do not agree in distribution, although they are related by a similarity transform. Hence, their eigenvalues are the same, but their eigenvectors will not agree. It is the eigenvectors of $\mathbf{C}$ that are of interest to us: they are the principal components. The tl;dr is that the reduction to only diagonal spiked covariances is fine when looking only at the eigenvalues; when information about the actual principal components is needed at the end, the original full (not necessarily diagonal) spiked covariance is needed.

3.2. Bulk Eigenvalue Distribution of Spiked Covariance Models

The last chapter was devoted to understanding the behavior of the (random) eigenvalues of a Wigner matrix $\mathbf{X}^\top\mathbf{X}$ where $\mathbf{X} \in \mathbb{M}_{m\times N}$ has i.i.d. entries (with mean 0 and variance 1) and $m, N \to \infty$ such that $m/N \to \varrho > 0$. Following Corollary 3.3 and the preceding discussion, we now return to denoting by $\mathbf{X}$ an $m \times N$ feature matrix that is pure noise (with all i.i.d. entires of mean 0 and variance 1); i.e. we dispense with $\mathbf{Z}$ (in the Gaussian case) and again call it $\mathbf{X}$. We want to understand more general "Wishart" matrices corresponding to spiked covariance models, of the form

$$\mathbf{W}_C = \frac{1}{N}\mathbf{X}^\top C\mathbf{X}$$

where $C = I_m + D_r$ is a low-rank diagonal perturbation of the identity matrix, cf. (3.1). As in Section 2.4, to understand the bulk behavior of the eigenvalues, it will suffice to compute their

sample moments (asymptotically), and therefore, as before, we wish to compute the (expected) *matrix moments*

$$\mathrm{tr}_N[\mathbf{W}_C^k], \qquad k \in \mathbb{N}.$$

As a warm-up, we begin with the case $k = 1$. Note that

$$\mathrm{tr}_N(\mathbf{W}_C) = \frac{1}{N^2}\sum_{i=1}^{N}[\mathbf{X}^\top C\mathbf{X}]_{ii} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{a,b=1}^{m}[\mathbf{X}^\top]_{ia}[C]_{ab}[\mathbf{X}]_{bi}. \tag{3.4}$$

Fortunately, C is a diagonal matrix, so $[C]_{ab} = 0$ if $a \neq b$. Thus

$$\mathrm{tr}_N(\mathbf{W}_C) = \frac{1}{N^2}\sum_{a=1}^{m}[C]_{aa}\sum_{i=1}^{N}[\mathbf{X}^\top]_{ia}[\mathbf{X}]_{ai} = \frac{1}{N^2}\sum_{a=1}^{m}[C]_{aa}\sum_{i=1}^{N}[\mathbf{X}]_{ai}^2.$$

This is close to the expression we found for the trace of the first moment of a Wishart matrix, in Example 2.18. Indeed, let's compare (denoting the standard Wishart matrix as $\mathbf{W}_I$ since it is the "spiked" Wishart matrix with covariance $C = I$):

$$\begin{aligned}\mathrm{tr}_N(\mathbf{W}_C) - \mathrm{tr}_N(\mathbf{W}_I) &= \frac{1}{N^2}\sum_{a=1}^{m}[C]_{aa}\sum_{i=1}^{N}[\mathbf{X}]_{ai}^2 - \frac{1}{N^2}\sum_{a=1}^{m}\sum_{i=1}^{N}[\mathbf{X}]_{ai}^2 \\ &= \frac{1}{N^2}\sum_{a=1}^{m}[C - I_m]_{aa}\sum_{i=1}^{N}[\mathbf{X}]_{ai}^2.\end{aligned}$$

Now, $C = I_m + D_r$, and so $[C - I_m]_{aa} = [D_r]_{aa}$; this is 0 whenever $a > r$. We have

$$\mathrm{tr}_N(\mathbf{W}_C) - \mathrm{tr}_N(\mathbf{W}_I) = \frac{1}{N^2}\sum_{a=1}^{r}\gamma_a\sum_{i=1}^{N}[\mathbf{X}]_{ai}^2.$$

To understand the asymptotic behavior of this random variable, we begin with its expectation:

$$\mathbb{E}\left[\,\mathrm{tr}_N(\mathbf{W}_C) - \mathrm{tr}_N(\mathbf{W}_I)\right] = \frac{1}{N^2}\sum_{a=1}^{r}\gamma_a\sum_{i=1}^{N}\mathbb{E}\left([\mathbf{X}]_{ai}^2\right) = \frac{1}{N}\sum_{a=1}^{r}\gamma_a$$

since $\mathbb{E}\left([\mathbf{X}]_{ai}^2\right) = 1$ for all (a, i). Since r is fixed, not varying with N, $\sum_{a=1}^{r}\gamma_a$ is just some positive constant, and so $\mathbb{E}\left[\,\mathrm{tr}_N(\mathbf{W}_C) - \mathrm{tr}_N(\mathbf{W}_I)\right] \to 0$ as $N \to \infty$ (regardless of the behavior of m, in fact). Actually, not just the expected value, but the actual random variable $\mathrm{tr}_N(\mathbf{W}_C) - \mathrm{tr}_N(\mathbf{W}_I)$ converges to 0 as $N \to \infty$; this can be seen by computing the variance. Indeed, all of the random variables $\gamma_a[\mathbf{X}]_{ai}^2$ are independent (over all choices of (a, i)), and therefore

$$\mathrm{Var}\left(\,\mathrm{tr}_N(\mathbf{W}_C) - \mathrm{tr}_N(\mathbf{W}_I)\right) = \frac{1}{N^4}\sum_{a=1}^{r}\sum_{i=1}^{N}\mathrm{Var}(\gamma_a[\mathbf{X}]_{ai}^2).$$

The $\frac{1}{N^4}$ comes from the relation $\mathrm{Var}(aY) = a^2\mathrm{Var}(Y)$. For the same reason, $\mathrm{Var}(\gamma_a[\mathbf{X}]_{ai}^2) = \gamma_a^2\mathrm{Var}([\mathbf{X}]_{ai}^2)$. As the $[\mathbf{X}]_{ai}$ are all identically distributed, $\mathrm{Var}([\mathbf{X}]_{ai}^2) = \mathrm{Var}([\mathbf{X}]_{11}^2)$ is constant (in the usual model we're employing where these are $\mathcal{N}(0, 1)$ random variables, the constant is 3). Thus

$$\mathrm{Var}\left(\,\mathrm{tr}_N(\mathbf{W}_C) - \mathrm{tr}_N(\mathbf{W}_I)\right) = \frac{\mathrm{Var}([\mathbf{X}]_{11}^2)}{N^4}\sum_{a=1}^{r}\sum_{i=1}^{N}\gamma_a^2 = \frac{\mathrm{Var}([\mathbf{X}]_{11}^2)}{N^3}\sum_{a=1}^{r}\gamma_a^2.$$

Again, since r is fixed, independent of N, this tends to 0 as $N \to \infty$. Thus

$$\lim_{N\to\infty} [\operatorname{tr}_N(\mathbf{W}_C) - \operatorname{tr}_N(\mathbf{W}_I)] = 0.$$

(The limit is in probability; however, since the variance is actually $O(1/N^3)$, by the Borel–Cantelli lemma, it is also almost sure.) As we showed in Example 2.18, $\operatorname{tr}_N(\mathbf{W}_I)$ converges to $\varrho = \lim \frac{m}{N}$, and so we conclude that also $\operatorname{tr}_N(\mathbf{W}_C) \to \varrho$ in this scaling regime.

REMARK 3.5. We didn't even have to assume r is fixed for this calculation to work. In fact, we could perfectly well have $r = m$, with a full-rank perturbation, provided that the two sums

$$\sum_{a=1}^{m} \gamma_a, \qquad \sum_{a=1}^{m} \gamma_a^2$$

are both uniformly bounded as m grows. As we proceed in the sequel, we would need higher powers and more complicated sums for this to work, so we will stick to the low-rank situation where r does not grow with N and m.

Following this line of reasoning, we can approach the higher moments similarly.

THEOREM 3.6. *Let $k \in \mathbb{N}$, and let $m, N \to \infty$ such that $m/N \to \varrho > 0$. Let $C = I_m + D_r$ be a low-rank diagonal perturbation of the identity covariance. Then*

$$\lim_{N\to\infty} [\operatorname{tr}_N(\mathbf{W}_C^k) - \operatorname{tr}_N(\mathbf{W}_I^k)] = 0.$$

PROOF. We begin by expanding $\operatorname{tr}_N[\mathbf{W}_C^k]$ exactly as in (2.19):

$$\operatorname{tr}_N[\mathbf{W}_C^k] = \frac{1}{N} \sum_{i_1,\dots,i_k=1}^{N} [\mathbf{W}_C]_{i_1 i_2} [\mathbf{W}_C]_{i_2 i_3} \cdots [\mathbf{W}_C]_{i_{k-1} i_k} [\mathbf{W}_C]_{i_k i_1}.$$

Now, $\mathbf{W}_C = \frac{1}{N}\mathbf{X}^\top C \mathbf{X}$ where C is diagonal, and so following the inside sum in (3.4), we have

$$[\mathbf{W}_C]_{ij} = \frac{1}{N} \sum_{a=1}^{N} [\mathbf{X}]_{ai} [C]_{aa} [\mathbf{X}]_{aj}.$$

Hence

$$\begin{aligned}
\operatorname{tr}_N[\mathbf{W}_C^k] &= \frac{1}{N^{k+1}} \sum_{i_1,\dots,i_k=1}^{N} \sum_{a_1,\dots,a_k=1}^{m} [\mathbf{X}]_{a_1 i_1} [C]_{a_1 a_1} [\mathbf{X}]_{a_1 i_2} \cdots [\mathbf{X}]_{a_k i_k} [C]_{i_k i_k} [\mathbf{X}]_{a_k i_1} \\
&= \frac{1}{N^{k+1}} \sum_{a_1,\dots,a_k=1}^{m} [C]_{a_1 a_1} \cdots [C]_{a_k a_k} \sum_{i_1,\dots,i_k=1}^{N} [\mathbf{X}]_{a_1 i_1} [\mathbf{X}]_{a_1 i_2} \cdots [\mathbf{X}]_{a_k i_k} [\mathbf{X}]_{a_k i_1}.
\end{aligned}$$

Comparing this to (2.20),

$$\operatorname{tr}_N(\mathbf{W}_I^k) = \frac{1}{N^{k+1}} \sum_{i_1,\dots,i_k=1}^{N} \sum_{a_1,a_2,\dots,a_k=1}^{m} [\mathbf{X}]_{a_1 i_1} [\mathbf{X}]_{a_1 i_2} [\mathbf{X}]_{a_2 i_2} [\mathbf{X}]_{a_2 i_3} \cdots [\mathbf{X}]_{a_k i_k} [\mathbf{X}]_{a_k i_1}$$

we see that the sums are closely aligned. Subtracting them, we have

$$\begin{aligned}
&\operatorname{tr}_N[\mathbf{W}_C^k] - \operatorname{tr}_N[\mathbf{W}_I^k] \\
&\quad = \frac{1}{N^{k+1}} \sum_{a_1,\dots,a_k=1}^{m} ([C]_{a_1 a_1} \cdots [C]_{a_k a_k} - 1) \sum_{i_1,\dots,i_k=1}^{N} [\mathbf{X}]_{a_1 i_1} [\mathbf{X}]_{a_1 i_2} \cdots [\mathbf{X}]_{a_k i_k} [\mathbf{X}]_{a_k i_1}.
\end{aligned}$$

Now, $C = I_m + D_r$, and so $[C]_{aa} = 1$ when $a > r$. Hence, in the sum over $a_1, \dots, a_k$, when all of the $a_j > r$, the product $[C]_{a_1 a_1} \cdots [C]_{a_k a_k}$ is equal to 1. This means the sum only goes up to r. We can also substitute in the values $[C]_{aa} = 1 + \gamma_a$ for $a \le r$:

$$\begin{aligned}&\operatorname{tr}_N[\mathbf{W}_C^k] - \operatorname{tr}_N[\mathbf{W}_I^k] \\ &\quad = \frac{1}{N^{k+1}} \sum_{a_1,\dots,a_k=1}^{r} ((1+\gamma_{a_1}) \cdots (1+\gamma_{a_k}) - 1) \sum_{i_1,\dots,i_k=1}^{N} [\mathbf{X}]_{a_1 i_1} [\mathbf{X}]_{a_1 i_2} \cdots [\mathbf{X}]_{a_k i_k} [\mathbf{X}]_{a_k i_1}.\end{aligned}$$

Taking expected values,

$$\begin{aligned}&\mathbb{E}\left(\operatorname{tr}_N[\mathbf{W}_C^k] - \operatorname{tr}_N[\mathbf{W}_I^k]\right) \\ &\quad = \frac{1}{N^{k+1}} \sum_{a_1,\dots,a_k=1}^{r} ((1+\gamma_{a_1}) \cdots (1+\gamma_{a_k}) - 1) \sum_{i_1,\dots,i_k=1}^{N} \mathbb{E}\left([\mathbf{X}]_{a_1 i_1} [\mathbf{X}]_{a_1 i_2} \cdots [\mathbf{X}]_{a_k i_k} [\mathbf{X}]_{a_k i_1}\right).\end{aligned}$$

The expected value of the product factors as a product of moments of the i.i.d. entries. The story of exactly which factorizations correspond to which indices involves partitions and gets complicated as we know; however, for our purposes, we only need an estimate. To simplify matters slightly, we make one assumption here: the the density of the entries $[\mathbf{X}]_{ai}$ is *symmetric*. (If we stick with the $\mathcal{N}(0,1)$ assumption, this holds true.) That means that all *odd* moments are 0, while the even moments (being expected values of an even power) are non-negative. Consequently, *all the terms in the above sum are non-negative*. Additionally, since each $\gamma_a > 0$, the terms $(1+\gamma_{a_1}) \cdots (1+\gamma_{a_k}) - 1$ are all positive as well. There are a fixed, finite number of such terms, and so they have a global maximum

$$M = \max\{(1+\gamma_{a_1}) \cdots (1+\gamma_{a_k}) - 1 \colon 1 \le a_1, \dots, a_k \le r\}.$$

Thus, we can make the estimate

$$\mathbb{E}\left(\operatorname{tr}_N[\mathbf{W}_C^k] - \operatorname{tr}_N[\mathbf{W}_I^k]\right) \le M \cdot \frac{1}{N^{k+1}} \sum_{a_1,\dots,a_k=1}^{r} \sum_{i_1,\dots,i_k=1}^{N} \mathbb{E}\left([\mathbf{X}]_{a_1 i_1} [\mathbf{X}]_{a_1 i_2} \cdots [\mathbf{X}]_{a_k i_k} [\mathbf{X}]_{a_k i_1}\right).$$

Compare the expression on the right-hand-side with (2.20); it is exactly equal to the kth moment of the Wishart ensemble, except now the dimension is r instead of m. We can now appeal to the result of Theorem 2.32: this moment converges to a certain degree-k polynomial in ϱ, where ϱ is the asymptotic ratio of the dimension (in this case r) to the number of samples N. But now we are holding r fixed as $N \to \infty$, so $\varrho = r/N \to 0$. Note that the kth moment polynomial in ϱ has no ϱ^0 term; hence, the whole things goes to 0 as $N \to \infty$.

This shows convergence in expectation; the full convergence in probability (or almost surely) is then completed by showing the variance goes to 0. The argument here is similar to the one needed for the variance of $\operatorname{tr}_N(\mathbf{W}_I^k)$, which we omitted in Section 2.4, and we also omit here. □

REMARK 3.7. The assumption of symmetric distribution for the entries of $\mathbf{X}$ is not actually needed for Theorem 3.6; it was just a convenient way to make the proof slick at the end, comparing to a Wishart ensemble based on a feature matrix of size $r \times N$. A more general argument can be applied using *Hölder's inequality*, which says that for any product of k random variables,

$$|\mathbb{E}(Y_1 Y_2 \cdots Y_k)| \le \left(\mathbb{E}(|Y_1|^k) \cdots \mathbb{E}(|Y_k|^k)\right)^{1/k}.$$

This can be applied to the terms above to factor out a constant no bigger than $\mathbb{E}(|[\mathbf{X}]_{11}|^k)$, showing that $|\mathbb{E}\left(\operatorname{tr}_N[\mathbf{W}_C^k] - \operatorname{tr}_N[\mathbf{W}_I^k]\right)|$ is

$$\leq \mathbb{E}(|[\mathbf{X}]_{11}|^k) \cdot \frac{1}{N^{k+1}} \sum_{a_1,\dots,a_k=1}^{r} ((1+\gamma_{a_1})\cdots(1+\gamma_{a_k}) - 1) \cdot N^k = O(\tfrac{1}{N}).$$

Furthermore, from this argument, we see that we don't actually need r to be fixes as $N, m \to \infty$; in fact, we can allow $r = m$ if we like, provided the γ_a decay as a grows fast enough that the above sums are uniformly bounded as m grows.

Theorem 3.6 shows that, for any spiked covariance model with covariance C, the spiked Wishart matrix $\mathbf{W}_C$ has the same asymptotic matrix moments as a standard Wishart matrix $\mathbf{W}_I$. Since the asymptotic matrix moments determine the asymptotic density of eigenvalues, we now conclude from Theorem 2.47 that all spiked covariance models have bulk eigenvalue distribution given by the Marčenko–Pastur distribution.

COROLLARY 3.8. *Let $\varrho > 0$, and let $(m_N)_{N\in\mathbb{N}}$ be a sequence of positive integers satisfying $\lim_{N\to\infty} \frac{m_N}{N} = \varrho$. Let $r \in \mathbb{N}$ be a fixed rank, let $C = I_m + D_r$ be a rank-r diagonal perturbation of the identity. Let $\mathbf{X} \in \mathbb{M}_{m_N \times N}$ be an i.i.d. feature matrix with entries of mean 0 and variance 1, and let $\mathbf{W}_C = \frac{1}{N}\mathbf{X}^\top C\mathbf{X}$ be the corresponding spiked Wishart matrix, with eigenvalue ensemble $(\boldsymbol{\lambda}^{(N)})$. Then for any $a < b$,*

$$\lim_{N\to\infty} h_{[a,b]}(\boldsymbol{\lambda}^{(N)}) = \frac{1}{N}\#\{j\colon \lambda_j^{(N)} \in [a,b]\} = \begin{cases} \displaystyle\int_a^b f_\varrho(x)\,dx, & \text{if } \varrho \geq 1 \\ (1-\varrho)\mathbb{1}_{[a,b]}(0) + \displaystyle\int_a^b f_\varrho(x)\,dx, & \text{if } \varrho \in (0,1) \end{cases}$$

where f_ϱ is the Marčenko–Pastur density (2.39).

That is: as far as the limit histogram of eigenvalues is concerned, spiked models made *no difference*. This might appear quite disappointing, and suggest there is something wrong with our model. But don't fret. Recall the discussion surrounding (2.42) and Bai–Yin's Theorem 2.48. The limit of the *proportion* of eigenvalues in a given interval tells us nothing about the *extreme* eigenvalues. Bai and Yin proved that, in the non-spiked case $(C = I)$, the largest eigenvalue sticks to the edge of the Marčenko–Pastur density. As we will see next, this fails for spiked covariance models: they have **outlier eigenvalues**.