EIGENVECTORS OF THE CORRELATION MATRIX

NICK STILES

1. Introduction

Suppose we have N observable random variables $X_1, ..., X_N$ with a correlation matrix given by $\mathbf{C}$ such that $\mathbf{C}_{ij}$ is the correlation between X_i and X_j. In general, comovements of the random variables are determined by a few ($r << N$) latent random variables which can be represented as uncorrelated linear combinations of the X_i's. When performing Principal Component Analysis (PCA), we determine these linear combinations as the eigenvectors corresponding to the highest eigenvalues of $\mathbf{C}$. This method of analyzing the spectral decomposition of $\mathbf{C}$ is common in many applications, such as medicine, genomics, physics, machine learning, image processing, finance, and engineering.

Unfortunately, we don't know $\mathbf{C}$, as we only have samples from the X_i's. Assuming we have T samples, we can form the empirical correlation matrix $\mathbf{E}$ such that $\mathbf{E}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}_{jt}$, or $\mathbf{E} = \frac{1}{T}\mathbf{X}\mathbf{X}^T$, where $\mathbf{X} \in \mathbb{R}^{N\times T}$ is the (normalized) matrix of observations. Of course, $\mathbf{E}$ contains the sample correlations rather than the true ones, so its eigenvectors will differ from those of $\mathbf{C}$. The goal of this project is to study how "close" the eigenvectors of $\mathbf{E}$ are to those of $\mathbf{C}$. We conduct a literature review, borrowing heavily from another literature review [2].

In section 2, we discuss various approaches to solving this problem. In section 3, we present some of the important tools from random matrix theory which are necessary to prove our results. In section 4, we state and prove our main results. In section 5, we discuss the results and conclude.

2. Framing the Mathematical Problem

To find an answer to our central question, we must first make decisions about the setup of the problem. The first question one might ask is whether we should pursue deterministic or random results. In the deterministic case, we would examine the eigenvectors of $\mathbf{C}+\mathbf{N}$ or $\sqrt{\mathbf{C}}\mathbf{N}\sqrt{\mathbf{C}}$ for some structured "perturbation" $\mathbf{N}$. For example, [3, 4, 5] all examine the problem from this perspective. One of the advantages of this approach is that we can find stronger bounds: for example, in the references above, they find ℓ_∞ (uniform entrywise) bounds, which are stronger than the ℓ_2 bounds found in this paper. Also, this approach allows us to accommodate fat-tailed distributions of the X_i, which otherwise may cause our probabilistic machinery to break down.

Unfortunately, these results often require many assumptions on the structure of $\mathbf{C}$ and $\mathbf{N}$, which may be hard to guarantee given that we only know $\mathbf{E}$. In the probabilistic approach, we let $\mathbf{E} = \sqrt{\mathbf{C}}\mathbf{N}\sqrt{\mathbf{C}}$ for a random matrix $\mathbf{N}$ and compute

Date: December 15, 2023.

expected "overlaps" between the eigenvectors of $\mathbf{E}$ and those of $\mathbf{C}$. Still, however, we must decide how to view the probabilistic problem. Specifically, we can either pursue results for fixed T, N, or we can pursue asymptotic results. In the former case, John Wishart [9] derived the exact density of $\frac{1}{T}\mathbf{X}\mathbf{X}^T$ for any N, T in the case that $\mathbf{X}$ has Gaussian entries with covariance $\mathbf{C}$. On the other hand, we would like results for more general distributions. Hence, the tools of infinite random matrix theory are helpful if we can send N, T or both to infinity. This assumption is justified in many applications in which our data matrix $\mathbf{X}$ is high-dimensional. If we fix N and sent $T \to \infty$, random matrix theory isn't necessary. As we are fixing the number of parameters and are given infinite observations, the law of large numbers tells us that in this case $\mathbf{E} \to \mathbf{C}$ almost surely. On the other hand, in many large datasets, N is approximately on the order of T. For example, suppose we have a set of 500 stocks whose daily prices are observed over 10 years. Then $N = 500, T = 2500$. We are thus justified to examine the case where $N, T \to \infty$ at some fixed asymptotic ratio $N/T = q = \Theta(1)$. This is the case examined in [2], and the one we explore for the remainder of the project.

3. Tools from Free Probability & Random Matrix Theory

In this section, we will define some of the important transforms and laws used in the proofs in our main results. We will also show how the problem at hand can be formulated as a problem about free products.

First, suppose we can write $\mathbf{X} = \sqrt{\mathbf{C}}\mathbf{Y}$ for some random matrix $\mathbf{Y}$ such that $\mathbf{E}[\mathbf{Y}_{it}] = \mathbf{E}[\mathbf{Y}_{it}] = 0, \mathbf{E}[\mathbf{Y}_{it}^2] = 1$, and $\mathbf{E}[\mathbf{Y}_{it}^4]$ is bounded. This assumption essentially requires that the distribution of the X_i is not too fat-tailed. Then let us set $\mathbf{W} = \frac{1}{T}\mathbf{Y}\mathbf{Y}^T$, so that

$$\mathbf{E} = \frac{1}{T}\mathbf{X}\mathbf{X}^T = \frac{1}{T}\sqrt{\mathbf{C}}\mathbf{Y}\mathbf{Y}^T\sqrt{\mathbf{C}} = \sqrt{\mathbf{C}}\mathbf{W}\sqrt{\mathbf{C}}.$$

Before proceeding, we list a couple of definitions and a remark.

Definition 3.1 (Wishart Matrix)**.** A *Wishart matrix* is $\frac{1}{T}\mathbf{G}\mathbf{G}^T$ for $\mathbf{G} \in \mathbb{R}^{N\times T}$ with i.i.d. $N(0, 1)$ entries.

Remark 3.2. $\mathbf{W}$ as defined above is a Wishart matrix in the limit as $N, T \to \infty$.

Definition 3.3 (Free Product)**.** The *free product* of two deterministic matrices $\mathbf{A}$ and $\mathbf{B}$ is the random matrix

$$\sqrt{\mathbf{A}}\mathbf{Q}\mathbf{B}\mathbf{Q}^T\sqrt{\mathbf{A}}.$$

Equivalently, if $\mathbf{B}$ is a random matrix with an orthogonally invariant distribution, i.e., $\mathbf{B}$ has the same distribution as $\mathbf{Q}\mathbf{B}$ for any fixed orthogonal matrix $\mathbf{Q}$, then the free product of $\mathbf{A}$ and $\mathbf{B}$ is the random matrix

$$\sqrt{\mathbf{A}}\mathbf{B}\sqrt{\mathbf{A}}.$$

Now the key observation is that, with $\mathbf{W}$ as defined above, we have that $\mathbf{E} = \sqrt{\mathbf{C}}\mathbf{W}\sqrt{\mathbf{C}}$ is the free product of $\mathbf{C}$ and the Wishart matrix $\mathbf{W}$. We learned in class that the R-transform of a free sum is the sum of the R-transforms of each of the random variables. The analogous transformation for free products is

the S-transform. Before defining it, let us first define the resolvent, the Cauchy transform and T-transform.

Definition 3.4 (Resolvent)**.** We define the *resolvent* of a random matrix $\mathbf{M}$ as

$$G_{\mathbf{M}}(z) = (z\mathbf{I} - \mathbf{M})^{-1}.$$

Definition 3.5 (Cauchy transform)**.** We define the *Cauchy transform* of a density function $f(x)$ of the eigenvalues of a random matrix $\mathbf{M}$ as

$$g_{\mathbf{M}}(z) = \int \frac{f(x)}{z - x} dx.$$

Remark 3.6. The asymptotic normalized trace of the resolvent $G_{\mathbf{M}}(z)$ as $N \to \infty$ is the Cauchy transform $g_{\mathbf{M}}(z)$.

Definition 3.7 (T-transform)**.** We define the T-transform as

$$T_{\mathbf{M}}(z) = z g_{\mathbf{M}}(z) - 1.$$

Definition 3.8. We define the S-transform of a random matrix $\mathbf{M}$ as

$$S_{\mathbf{M}}(\omega) = \frac{\omega + 1}{\omega T_{\mathbf{M}}^{-1}(\omega)}.$$

Proposition 3.9. *If* $\mathbf{M}$ *is the free product of* $\mathbf{A}$ *and* $\mathbf{B}$*, then*

$$S_{\mathbf{M}}(\omega) = S_{\mathbf{A}}(\omega) S_{\mathbf{B}}(\omega).$$

Thus, we have a tool, the S-transform, which we can use to relate $\mathbf{E}$ and $\mathbf{C}$. Specifically, we make use of the fact that the S-transform of $\mathbf{W}$ is

$$S_{\mathbf{W}}(\omega) = \frac{1}{1 + q\omega},$$

which follows from the proof of the Marcenko-Pastur law.

4. Main Results

4.1. **Defining overlap.** We first must define a metric for overlap of eigenvectors. Let the spectral decompositions of $\mathbf{E}$ and $\mathbf{C}$ be

$$\mathbf{E} = \sum_{i=1}^{N} \mu_i \mathbf{u}_i \mathbf{u}_i^T \qquad\qquad \mathbf{C} = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$$

where $\lambda_1 > ... > \lambda_N$, $\mu_1 > ... > \mu_N$ and all eigenvectors are normalized to length 1. Define the overlap as

$$\Phi(\mu_i, \lambda_j) = N\mathbb{E}[\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2],$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^N$ such that $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v}$. We put the square inside the expectation to account for the fact that eigenvectors are only defined up to change in sign. Our goal will be to derive conclusions about $\Phi(\mu_i, \lambda_j)$.

4.2. **Spiked covariance model.** We now present the "spiked covariance model", which allows us to distinguish between *bulk* eigenvectors/values and *outlier* eigenvectors/values. In this model, the eigenvalue distribution of $\mathbf{E}$ consists of a "bulk" of smaller eigenvalues and a finite number r of "spikes" at higher values, representing the positive outlier eigenvalues. We deal with this phenomenon by studying the *spikeless covariance matrix*

$$\underline{\mathbf{C}} = \sum_{i=1}^{N} \underline{\lambda_i} \mathbf{v}_i \mathbf{v}_i^T \text{ where } \underline{\lambda_i} = \begin{cases} \lambda_{r+1} & \text{if } i \leq r \\ \lambda_i & \text{if } i > r, \end{cases}$$

which coerces the top r eigenvalues of $\mathbf{C}$ down to λ_{r+1}. We then can form

$$\underline{\mathbf{E}} = \sqrt{\underline{\mathbf{C}}} \mathbf{W} \sqrt{\underline{\mathbf{C}}},$$

which is the spikeless empirical covariance matrix.

4.3. **Bulk eigenvectors.** Our first theorem tells us the expected overlap between any bulk eigenvector from the sample covariance matrix with any (bulk or outlier) eigenvector from the true covariance matrix.

Theorem 4.1. *For all $i > r$ and for all $j = 1, ..., N$, we have*

$$\Phi(\mu_i, \lambda_j) = \frac{q \mu_i \lambda_j}{(\lambda_j(1-q) - \mu_i + q\mu_i\lambda_j \mathrm{Re}(g_{\mathbf{E}}(\mu_i)))^2 + q^2 \mu_i^2 \lambda_j^2 \pi^2 f_{\mathbf{E}}(\mu_i)^2},$$

which is $O(1)$ whenever $q > 0$.

Thus, sample eigenvectors in the bulk (i.e., $\mathbf{u}_i$ for $i > r$) are "delocalized" in the population basis. This implies sample eigenvectors corresponding to smaller eigenvalues contain little or no information about population eigenvectors. The proof of Theorem 4.1 is quite involved, but here we will give a sketch which relies on some results we leave unproven. We give references to proofs of all unproven results.

Sketch of Proof of Theorem 4.1. We begin with the proof by using an equation proven in section 2.4 of [2], namely that

$$z G_{\mathbf{E}}(z)_{ij} = Z(z) G_{\mathbf{C}}(Z(z))_{ij}, \text{ where } Z(z) := z S_{\mathbf{W}}(z g_{\mathbf{E}}(z) - 1). \tag{4.1}$$

Using the fact given above that $S_{\mathbf{W}}(\omega) = 1/(1 + q\omega)$, we have

$$Z(z) = \frac{1}{1 - q + q z g_{\mathbf{E}}(z)}. \tag{4.2}$$

Now, recall the inversion formula from class stating

$$f(x) = \lim_{\epsilon \to 0^+} \mathrm{Im}(g(x - i\epsilon)).$$

where g is the Cauchy transform of f. We will derive an inversion formula similar to this for the full resolvent G. From the definition of the resolvent, we have

$$G_{\mathbf{E}}(z) = \sum_{i=1}^{N} \frac{\mathbf{u}_i \mathbf{u}_i^T}{z - \mu_i}.$$

Noting that the eigenvectors $\mathbf{v}_j$ of $\mathbf{C}$ are deterministic, we have for any $j = 1, ..., n$

$$\langle \mathbf{v}_j, G_{\mathbf{E}}(z)\mathbf{v}_j \rangle = \sum_{i=1}^{N} \frac{\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2}{z - \mu_i}.$$

Next, we will take the limit of the RHS as $N \to \infty$. To evaluate it, we use theorem 3.12 in [6], which states that the eigenvalues of $\mathbf{E}$ converge to their quantile positions:

$$\mu_i \to \gamma_i, \text{ where } \frac{i}{N} = \int_{-\infty}^{\gamma_i} f_{\mathbf{E}}(\mu) d\mu.$$

Thus, taking the limit as $N \to \infty$, we get

$$\langle \mathbf{v}_j, G_{\mathbf{E}}(z)\mathbf{v}_j \rangle = \int \frac{\Phi(\mu, \lambda_j) f_{\mathbf{E}}(\mu)}{z - \mu} d\mu,$$

where $\Phi(\mu, \lambda_j)$ is the "smoothed" overlap found by averaging $\Phi(\mu, \lambda_j)$ is found by averaging over a small interval $d\mu$ around μ. Setting $z = \mu_i - \mathrm{i}\epsilon$, where i is the complex number $\sqrt{-1}$. We then apply the Plemelj-Sokhotski theorem to attain

$$\Phi(\mu_i, \lambda_j) = \frac{1}{\pi f_{\mathbf{E}}(\mu_i)} \lim_{\epsilon \to 0^+} \mathrm{Im}(\langle \mathbf{v}_j, G_{\mathbf{E}}(\mu_i - \mathrm{i}\epsilon)\mathbf{v}_j \rangle).$$

If we rotate into a basis such that $\mathbf{C}$ is diagonal, we that our earlier result specializes to

$$G_{\mathbf{E}}(z)_{ij} = \frac{\delta_{ij}}{z - \mu_i(1 - q + q g_{\mathbf{E}}(z))}.$$

Plugging this back into our inversion formula, we finally obtain the desired result

$$\Phi(\mu_i, \lambda_j) = \frac{q\mu_i\lambda_j}{(\lambda_j(1-q) - \mu_i + q\mu_i\lambda_j \mathrm{Re}(g_{\mathbf{E}}(\mu_i)))^2 + q^2\mu_i^2\lambda_j^2\pi^2 f_{\mathbf{E}}(\mu_i)^2}.$$

In this final step, we hide a few algebraic steps which can be found in full detail in [7] in the proof of Theorem 1.3. □

4.4. **Outlier eigenvectors.** Our second theorem tells us the expected overlap between any outlier sample eigenvector with any (bulk or outlier) eigenvector from the true covariance matrix.

Theorem 4.2. *For all $i \leq r$ and for all $j = 1, ..., N$, we have*

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = \delta_{ij}\mu_i \frac{\theta'(\mu_i)}{\theta(\mu_i)} + O\left(\frac{1}{\sqrt{N}}\right),$$

where

$$\theta(\mu_i) = g_{\underline{\mathbf{S}}}^{-1}(1/\mu_i),$$

where $g_{\underline{\mathbf{S}}}^{-1}$ is the inverse Cauchy transform of the matrix $\underline{\mathbf{S}} = \mathbf{Y}^T\underline{\mathbf{C}}\mathbf{Y}$, which is the $T \times T$ "dual" of $\underline{\mathbf{E}}$.

Thus, outlier sample eigenvectors (i.e., $\mathbf{u}_i$ for $i \leq r$) are concentrated on the surface of a cone around the corresponding outlier population eigenvector $\mathbf{v}_i$. On the other hand, $\mathbf{u}_i$ is "delocalized" in any population eigendirection $\mathbf{v}_j$ for $j \neq i$. This implies sample eigenvectors corresponding to outlier eigenvalues contain

some information about the underlying population eigenvector, and furthermore carry no information about any other population eigenvector. We will again give a sketch of the proof which relies on some results we leave unproven.

Sketch of Proof of Theorem 4.2. An important property of outlier eigenvalues μ_i is that $f_{\mathbf{E}}(\mu_i) = 0$. This can be seen from the fact that $\mathbf{E}$ is a finite-rank perturbation of $\underline{\mathbf{E}}$ and that clearly μ_i is outside the spectrum of $\underline{\mathbf{E}}$, which only consists of the bulk. This immediately shows we cannot rely on many of the steps we took in the proof of Theorem 4.1.

Instead, we begin with the fact that $\mu_i \to \theta(\lambda_i)$, which is proven on page 51 in [2]. We then define r discs D_i for $i = 1, ..., r$ such that D_i is centered on $\theta(\lambda_i)$ with radius such that no disk D_i contains $\theta(\lambda_j)$ for $j \neq i$. Letting Γ_i be the boundaries of such discs, we have by Cauchy's integral formula and Cauchy's integral theorem that

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = \frac{1}{2\pi \mathrm{i}} \oint_{\Gamma_i} \frac{\langle \mathbf{u}_i \mathbf{v}_j \rangle^2}{z - \mu_i} dz = \frac{1}{2\pi \mathrm{i}} \sum_{k=1}^{N} \oint_{\Gamma_i} \frac{\langle \mathbf{u}_k, \mathbf{v}_j \rangle^2}{z - \mu_k} dz = \frac{1}{2\pi \mathrm{i}} \oint_{\Gamma_i} \langle \mathbf{v}_j, G_{\mathbf{E}}(z) \mathbf{v}_j \rangle dz. \tag{4.3}$$

Importantly, there is no expectation in this expression, because we are conditioning on an event which holds with high probability under our assumptions. (See proof of Lemma 4.24 in [1] for details.)

Unfortunately, the final integral is hard to compute, as $G_{\mathbf{E}}(z)$ blows up around $\theta(\lambda_i)$. Instead, we thus consider $G_{\underline{\mathbf{E}}}(z)$, which no longer blows up near $\theta(\lambda_i)$. We need to derive a formula for the projection of $G_{\mathbf{E}}(z)$ onto the population outlier eigenvectors in terms of $G_{\underline{\mathbf{E}}}(z)$. To do so, we define matrices $\mathbf{V}$ and $\mathbf{D}$ such that

$$\mathbf{C} = \underline{\mathbf{C}}(\mathbf{I}_N + \mathbf{V}\mathbf{D}\mathbf{V}^T),$$

i.e., $\mathbf{V} \in \mathbb{R}^{N \times r}$ contains the first r eigenvectors of $\mathbf{C}$ and $\mathbf{D} \in \mathbb{R}^{d \times d}$ is diagonal with entries $\mathbf{D}_{ii} = d_i = \lambda_i - \lambda_{r+1}$. Then we can apply the Schur complement formula to obtain the expression

$$\mathbf{V}^T G_{\mathbf{E}}(z) \mathbf{V} = -\frac{1}{z} \left(\mathbf{D}^{-1} - \frac{\sqrt{\mathbf{I}_r + \mathbf{D}}}{\mathbf{D}} (\mathbf{D}^{-1} + \mathbf{I}_r - z\mathbf{V}^T G_{\underline{\mathbf{E}}}(z) \mathbf{V}) \frac{\sqrt{\mathbf{I}_r + \mathbf{D}}}{\mathbf{D}} \right).$$

The proof of this identity can be found on page 60 of [2]. We then recall equations 4.1, 4.2 from the proof of Theorem 4.1, which, when combined with the above identity and 4.3, gives

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = -\frac{1}{2\pi \mathrm{i}} \oint_{\Gamma_i} \frac{1}{z} \left(\frac{1}{d_j} - \frac{1 + d_j}{d_j^2} \cdot \frac{1}{d_j^{-1} + 1 - z \langle \mathbf{v}_j, G_{\underline{\mathbf{E}}}(z) \mathbf{v}_j \rangle} \right) dz.$$

We then use a fact proven as Lemma 4.19 in [1], which states that μ_i is an eigenvalue of $\mathbf{E}$ and not $\underline{\mathbf{E}}$ only if

$$d_j(\mu_i \langle \mathbf{v}_j, G_{\underline{\mathbf{E}}}(\mu_i) \mathbf{v}_j \rangle - 1) = 1.$$

Plugging this into the above and applying Cauchy's theorem, we find our final result that

$$\langle \mathbf{u}_i, \mathbf{v}_j \rangle^2 = \delta_{ij} \mu_i \frac{\theta'(\mu_i)}{\theta(\mu_i)} + O\left(\frac{1}{\sqrt{N}}\right),$$

where the detailed steps to reach this final conclusion are given in the proof of Lemma 4.26 in [1]. □

5. Discussion

In some sense, the results are disappointing. If we use the eigenvectors of $\mathbf{E}$ to estimate those of $\mathbf{C}$, we can essentially throw away any sample eigenvectors in the bulk, as these carry no information about the underlying population eigenvectors. On the other hand, the results about outlier eigenvectors is encouraging: these eigenvectors do contain information about the corresponding population outlier eigenvectors. What's even more surprising is that we can compute, to a good approximation, the angle between these eigenvectors. This implies that we "know the unknowns", and allows us to proceed with greater caution when attempting to create estimators of $\mathbf{C}$. Indeed, these results add insight to the original conclusion found in [8] that "the best we can do" in estimating the eigenbasis of $\mathbf{C}$ is to simply stick with the default basis given to us by $\mathbf{E}$.

References

[1] Joel Bun. An optimal rotational invariant estimator for general covariance matrices: The outliers. 02 2018.

[2] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666:1–109, January 2017. URL: http://dx.doi.org/10.1016/j.physrep.2016.10.005, doi:10.1016/j.physrep.2016.10.005.

[3] Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*, vol 75:pg. 1–44, 2013.

[4] Jianqing Fan, Weichen Wang, and Yichao Zhong. Robust covariance estimation for approximate factor models. *Journal of Econometrics*, 2017.

[5] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An ∞ eigenvector perturbation bound and its application to robust covariance estimation. *arXiv:1603.03516v2*, Jun 2017.

[6] Antti Knowles and Jun Yin. Anisotropic local laws for random matrices, 2016. arXiv:1410.3516.

[7] Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles, 2009. arXiv:0911.3010.

[8] A. Takemura. An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population. Tech. rep., DTIC Document, 1983.

[9] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1-2):32–52, December 1928. doi:10.1093/biomet/20A.1-2.32.