How Close Are the Eigenvectors of the Sample and Actual Covariance Matrices?

Andreas Loukas [1]

Abstract

How many samples are sufficient to guarantee that the eigenvectors of the sample covariance matrix are close to those of the actual covariance matrix? For a wide family of distributions, including distributions with finite second moment and sub-gaussian distributions supported in a centered Euclidean ball, we prove that the inner product between eigenvectors of the sample and actual covariance matrices decreases proportionally to the respective eigenvalue distance and the number of samples. Our findings imply *non-asymptotic* concentration bounds for eigenvectors and eigenvalues and carry strong consequences for the non-asymptotic analysis of PCA and its applications. For instance, they provide conditions for separating components estimated from $O(1)$ samples and show that even few samples can be sufficient to perform dimensionality reduction, especially for low-rank covariances.

1 Introduction

The covariance matrix C of an n-dimensional distribution is an integral part of data analysis, with numerous occurrences in machine learning and signal processing. It is therefore crucial to understand how close is the *sample covariance*, i.e., the matrix $\widetilde{C}$ estimated from a finite number of samples m, to the actual covariance matrix. Following developments in the tools for the concentration of measure, (Vershynin, 2012) showed that a sample size of $m = O(n)$ is up to iterated logarithmic factors sufficient for all distributions with finite fourth moment supported in a centered Euclidean ball of radius $O(\sqrt{n})$. Similar results hold for sub-exponential (Adamczak et al., 2010) and finite second moment distributions (Rudelson, 1999).

We take an alternative standpoint and ask if we can do better when only a subset of the spectrum is of interest. Concretely, our objective is to characterize how many samples are sufficient to guarantee that an eigenvector and/or eigenvalue of the sample and actual covariance matrices are, respectively, sufficiently close. Our approach is motivated by the observation that methods that utilize the covariance commonly prioritize the estimation of principal eigenspaces. For instance, in (local) principal component analysis we are usually interested in the direction of the first few eigenvectors (Berkmann & Caelli, 1994; Kambhatla & Leen, 1997), where in linear dimensionality reduction one projects the data to the span of the first few eigenvectors (Jolliffe, 2002; Frostig et al., 2016). In the non-asymptotic regime, an analysis of these methods hinges on characterizing how close are the principal eigenvectors and eigenspaces of the sample and actual covariance matrices.

Our finding is that the "spectral leaking" occurring in the eigenvector estimation is strongly concentrated along the eigenvalue axis. In other words, the eigenvector $\widetilde{u}_i$ of the sample covariance is far less likely to lie in the span of an eigenvector u_j of the actual covariance when the eigenvalue distance $|\lambda_i - \lambda_j|$ is large, and the concentration of the distribution in the direction of u_j is small. This agrees with the intuition that principal components are easier to estimate, exactly because they are more likely to appear in the samples of the distribution.

We provide a mathematical argument confirming this phenomenon. Under fairly general conditions, we prove that

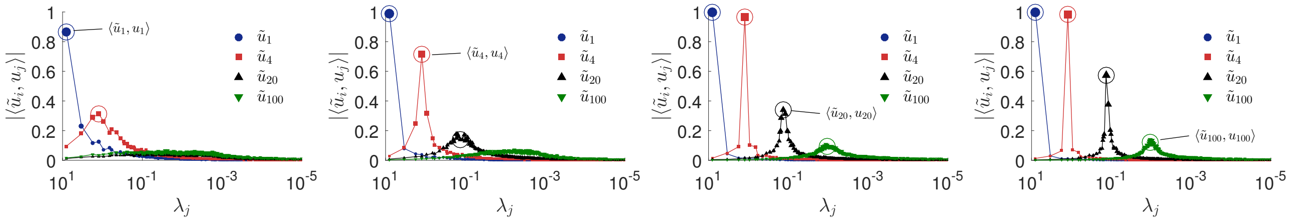
$$m = O\left(\frac{k_j^2}{(\lambda_i - \lambda_j)^2}\right) \quad \text{and} \quad m = O\left(\frac{k_i^2}{\lambda_i^2}\right) \tag{1}$$

samples are asymptotically almost surely (a.a.s.) sufficient to guarantee that $|\langle \widetilde{u}_i, u_j \rangle|$ and $|\widetilde{\lambda}_i - \lambda_i|/\lambda_i$, respectively, is small for all distributions with finite second moment. Here, k_j^2 is a measure of the concentration of the distribution in the direction of u_j. We also attain a high probability bound for sub-gaussian distributions supported in a centered Euclidean ball. Interestingly, our results lead to sample estimates for linear dimensionality reduction, and suggest that linear reduction is feasible even from few samples.

To the best of our knowledge, these are the first non-asymptotic results concerning the eigenvectors of the sam-

[1] École Polytechnique Fédérale de Lausanne, Switzerland. Correspondence to: Andreas Loukas <andreas.loukas@epfl.ch>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).



(a) $m = 10$ (b) $m = 100$ (c) $m = 500$ (d) $m = 1000$

Figure 1: Inner products $\langle \widetilde{u}_i, u_j \rangle$ are localized w.r.t. the eigenvalue axis. The phenomenon is shown for MNIST. Much fewer than $n = 784$ samples are needed to approximate u_1 and u_4.

ple covariance of non-Normal distributions. Previous studies have intensively investigated the limiting distribution of the eigenvalues of a sample covariance matrix (Silverstein & Bai, 1995; Bai, 1999), such as the smallest and largest eigenvalues (Bai & Yin, 1993) and the eigenvalue support (Bai & Silverstein, 1998). Eigenvectors and eigenprojections have attracted less attention; the main research thrust entails using tools from the theory of large-dimensional matrices to characterize limiting distributions (Anderson, 1963; Girko, 1996; Schott, 1997; Bai et al., 2007) and it has limited applicability in the non-asymptotic setting where the sample size m is small and n cannot be arbitrary large.

Differently, we use techniques from perturbation analysis and non-asymptotic concentration of measure. However, in contrast to arguments commonly used to reason about eigenspaces (Davis & Kahan, 1970; Yu et al., 2015; Huang et al., 2009; Hunter & Strohmer, 2010), our bounds can characterize weighted linear combinations of $\langle \widetilde{u}_i, u_j \rangle^2$ over i and j, and do not depend on the minimal eigenvalue gap separating two eigenspaces but rather on all eigenvalue differences. The latter renders them useful to many real datasets, where the eigenvalue gap is not significant but the eigenvalue magnitudes decrease sufficiently fast.

We also note two recent works targeting the non-assymptotic regime of Normal distributions. Shaghaghi and Vorobyov recently characterized the first two moments of the subspace projection error, a result which implies sample estimates (Shaghaghi & Vorobyov, 2015), but is restricted to specific projectors. A refined concentration analysis for spectral projectors of Normal distributions was also presented in (Koltchinskii & Lounici, 2015). Finally, we remark that there exist alternative estimators for the spectrum of the covariance with better asymptotic properties (Ahmed, 1998; Mestre, 2008). Instead, we here focus on the standard estimates, i.e., the eigenvalues and eigenvectors of the sample covariance.

2 Problem Statement and Main Results

Let $x \in \mathbb{C}^n$ be a sample of a multivariate distribution and denote by $x_1, x_2, \ldots, x_m$ the m independent samples used to form the sample covariance, defined as

$$\widetilde{C} = \sum_{p=1}^{m} \frac{(x_p - \bar{x})(x_p - \bar{x})^*}{m}, \tag{2}$$

where $\bar{x}$ is the sample mean. Denote by u_i the eigenvector of C associated with eigenvalue λ_i, and correspondingly for the eigenvectors $\widetilde{u}_i$ and eigenvalues $\widetilde{\lambda}_i$ of $\widetilde{C}$, such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. We ask:

Problem 1. *How many samples are sufficient to guarantee that the inner product $|\langle \widetilde{u}_i, u_j \rangle| = |\widetilde{u}_i^* u_j|$ and the eigenvalue gap $|\delta\lambda_i| = |\widetilde{\lambda}_i - \lambda_i|$ is smaller than some constant t with probability larger than ϵ?*

Clearly, when asking that all eigenvectors and eigenvalues of the sample and actual covariance matrices are close, we will require at least as many samples as needed to ensure that $\|\widetilde{C} - C\|_2 \leq t$. However, we might do better when only a subset of the spectrum is of interest. The reason is that inner products $|\langle \widetilde{u}_i, u_j \rangle|$ are strongly concentrated along the eigenvalue axis. To illustrate this phenomenon, let us consider the distribution constructed by the $n = 784$ pixel values of digit '1' in the MNIST database. Figure 1, compares the eigenvectors u_j of the covariance computed from all 6742 images, to the eigenvectors $\widetilde{u}_i$ of the sample covariance matrices $\widetilde{C}$ computed from a random subset of $m = 10$, 100, 500, and 1000 samples. For each $i = 1, 4, 20, 100$, we depict at λ_j the average of $|\langle \widetilde{u}_i, u_j \rangle|$ over 100 sampling draws. We observe that: (*i*) The magnitude of $\langle \widetilde{u}_i, u_j \rangle$ is inversely proportional to their eigenvalue gap $|\lambda_i - \lambda_j|$. (*ii*) Eigenvector $\widetilde{u}_j$ mostly lies in the span of eigenvectors u_j over which the distribution is concentrated.

We formalize these statements in two steps.

2.1 Perturbation arguments

First, we work in the setting of Hermitian matrices and notice the following inequality:

Theorem 3.2. *For Hermitian matrices C and $\widetilde{C} = \delta C + C$, with eigenvectors u_j and $\widetilde{u}_i$ respectively, the inequality*

$$|\langle \widetilde{u}_i, u_j \rangle| \leq \frac{2\,\|\delta C u_j\|_2}{|\lambda_i - \lambda_j|},$$

holds for $sgn(\lambda_i - \lambda_j)\, 2\widetilde{\lambda}_i > sgn(\lambda_i - \lambda_j)(\lambda_i + \lambda_j)$ *and* $\lambda_i \neq \lambda_j$.

The above stands out from standard eigenspace perturbation results, such as the sin(Θ) Theorem (Davis & Kahan, 1970) and its variants (Huang et al., 2009; Hunter & Strohmer, 2010; Yu et al., 2015) for three main reasons:

First, Theorem 3.2 characterizes the angle between *any* pair of eigenvectors allowing us to jointly bound any linear combination of inner-products. Though this often proves handy (c.f. Section 5), it is infeasible using sin(Θ)-type arguments. Second, classical bounds are not appropriate for a probabilistic analysis as they feature ratios of dependent random variables (corresponding to perturbation terms). In the analysis of spectral clustering, this complication was dealt with by assuming that $|\lambda_i - \lambda_j| \leq |\widetilde{\lambda}_i - \lambda_j|$ (Hunter & Strohmer, 2010). We weaken this condition at a cost of a multiplicative factor. In contrast to previous work, we also prove that the condition is met a.a.s. Third, previous bounds are expressed in terms of the minimal eigenvalue gap between eigenvectors lying in the interior and exterior of the subspace of interest. This is a limiting factor in practice as it renders the results only amenable to situations where there is a very large eigenvalue gap separating the subspaces. The proposed result improves upon this by considering every eigenvalue difference.

2.2 Concentration of measure

The second part of our analysis focuses on the covariance and has a statistical flavor. In particular, we give an answer to Problem 1 for various families of distributions.

In the context of *distributions with finite second moment*, we prove in Section 4.1 that:

Theorem 4.1. *For any two eigenvectors* $\widetilde{u}_i$ *and* u_j *of the sample and actual covariance respectively, and for any real number* $t > 0$*:*

$$\mathbf{P}\left(|\langle \widetilde{u}_i, u_j \rangle| \geq t\right) \leq \frac{1}{m} \left(\frac{2k_j}{t\,|\lambda_i - \lambda_j|} \right)^2, \qquad (3)$$

subject to the same conditions as Theorem 3.2.

For eigenvalues, we have the following corollary:

Corollary 2.1. *For any eigenvalues* λ_i *and* $\widetilde{\lambda}_i$ *of* C *and* $\widetilde{C}$*, respectively, and for any* $t > 0$*, we have*

$$\mathbf{P}\left(\frac{|\widetilde{\lambda}_i - \lambda_i|}{\lambda_i} \geq t \right) \leq \frac{1}{m} \left(\frac{k_i}{\lambda_i\, t} \right)^2 .$$

Term $k_j = (\mathbf{E}\left[\|xx^* u_j\|_2^2\right] - \lambda_j^2)^{1/2}$ captures the tendency of the distribution to fall in the span of u_j: the smaller the tail in the direction of u_j the less likely we are going to confuse $\widetilde{u}_i$ with u_j.

For *normal distributions*, we have that $k_j^2 = \lambda_j^2 + \lambda_j \mathrm{tr}(C)$ and the number of samples needed for $|\langle \widetilde{u}_i, u_j \rangle|$ to be small is $m = O(\mathrm{tr}(C)/\lambda_i^2)$ when $\lambda_j = O(1)$ and $m = O(\lambda_i^{-2})$ when $\lambda_j = O(\mathrm{tr}(C)^{-1})$. Thus for normal distributions, principal components u_i and u_j with $\min\{\lambda_i/\lambda_j, \lambda_i\} = \Omega(\mathrm{tr}(C)^{1/2})$ can be distinguished given a constant number of samples. On the other hand, estimating λ_i with small relative error requires $m = O(\mathrm{tr}(C)/\lambda_i)$ samples and can thus be achieved from very few samples when λ_i is large[1].

In Section 4.2, we also give a sharp bound for the family of distributions supported within a ball (i.e., $\|x\| \leq r$ a.s.).

Theorem 4.2. *For sub-gaussian distributions supported within a centered Euclidean ball of radius* r*, there exists an absolute constant* c *s.t. for any real number* $t > 0$*,*

$$\mathbf{P}\left(|\langle \widetilde{u}_i, u_j \rangle| \geq t\right) \leq exp\left(1 - \frac{c\, m\, \Phi_{ij}(t)^2}{\lambda_j\, \|x\|_{\Psi_2}^2} \right), \qquad (4)$$

where $\Phi_{ij}(t) = \frac{|\lambda_i - \lambda_j|\, t - 2\lambda_j}{2\,(r^2/\lambda_j - 1)^{1/2}} - 2\,\|x\|_{\Psi_2}$ *and subject to the same conditions as Theorem 3.2.*

Above, $\|x\|_{\Psi_2}$ is the sub-gaussian norm, for which we usually have $\|x\|_{\Psi_2} = O(1)$ (Vershynin, 2010). As such, the theorem implies that, whenever $\lambda_i \gg \lambda_j = O(1)$, the sample requirement is with high probability $m = O(r^2/\lambda_i^2)$.

These theorems solidify our experimental findings shown in Figure 1 and provide a concrete characterization of the relation between the spectrum of the sample and actual covariance matrix as a function of the number of samples, the eigenvalue gap, and the distribution properties. As exemplified in Section 5 for linear dimensionality reduction, we believe that our results carry strong implications for the non-asymptotic analysis of PCA-based methods.

3 Perturbation Arguments

Before focusing on the sample covariance matrix, it helps to study $\langle \widetilde{u}_i, u_j \rangle$ in the setting of Hermitian matrices. The presentation of the results is split in three parts. Section 3.1 starts by studying some basic properties of inner products of the form $\langle \widetilde{u}_i, u_j \rangle$, for any i and j. The results are used in Section 3.2 to provide a first bound on the angle between two eigenvectors, and refined in Section 3.3.

3.1 Basic observations

We start by noticing an exact relation between the angle of a perturbed eigenvector and the actual eigenvectors of C.

Lemma 3.1. *For every* i *and* j *in* $1, 2, \ldots, n$*, the relation* $(\widetilde{\lambda}_i - \lambda_j)\, (\widetilde{u}_i^* u_j) = \sum_{\ell=1}^{n} (\widetilde{u}_i^* u_\ell)\, (u_j^* \delta C u_\ell)$ *holds .*

[1] Though the same cannot be stated about the absolute error $|\delta\lambda_i|$, that is smaller for small λ_i.

Proof. The proof follows from a modification of a standard argument in perturbation theory. We start from the definition $\widetilde{C}\,\widetilde{u}_i = \widetilde{\lambda}_i\,\widetilde{u}_i$ and write

$$(C + \delta C)\,(u_i + \delta u_i) = (\lambda_i + \delta\lambda_i)\,(u_i + \delta u_i). \tag{5}$$

Expanded, the above expression becomes

$$\begin{aligned} C\delta u_i &+ \delta C u_i + \delta C \delta u_i \\ &= \lambda_i \delta u_i + \delta\lambda_i u_i + \delta\lambda_i \delta u_i, \end{aligned} \tag{6}$$

where we used the fact that $C u_i = \lambda_i u_i$ to eliminate two terms. To proceed, we substitute $\delta u_i = \sum_{j=1}^{n} \beta_{ij} u_j$, where $\beta_{ij} = \delta u_i^* u_j$, into (6) and multiply from the left by u_j^*, resulting to:

$$\begin{aligned} \sum_{\ell=1}^{n} \beta_{ij} u_j^* C u_\ell &+ u_j^* \delta C u_i + \sum_{\ell=1}^{n} \beta_{ij} u_j^* \delta C u_\ell \\ &= \lambda_i \sum_{\ell=1}^{n} \beta_{ij} u_j^* u_\ell + \delta\lambda_i u_j^* u_i + \delta\lambda_i \sum_{\ell=1}^{n} \beta_{ij} u_j^* u_\ell \end{aligned} \tag{7}$$

Cancelling the unnecessary terms and rearranging, we have

$$\begin{aligned} \delta\lambda_i u_j^* u_i &+ (\lambda_i + \delta\lambda_i - \lambda_j)\beta_{ij} \\ &= u_j^* \delta C u_i + \sum_{\ell=1}^{n} \beta_{ij} u_j^* \delta C u_\ell. \end{aligned} \tag{8}$$

At this point, we note that $(\lambda_i + \delta\lambda_i - \lambda_j) = \widetilde{\lambda}_i - \lambda_j$ and furthermore that $\beta_{ij} = \widetilde{u}_i^* u_j - u_i^* u_j$. With this in place, equation (8) becomes

$$\begin{aligned} \delta\lambda_i u_j^* u_i &+ (\widetilde{\lambda}_i - \lambda_j)\,(\widetilde{u}_i^* u_j - u_i^* u_j) \\ &= u_j^* \delta C u_i + \sum_{\ell=1}^{n} (\widetilde{u}_i^* u_\ell)\, u_j^* \delta C u_\ell - u_j^* \delta C u_i. \end{aligned} \tag{9}$$

The proof completes by noticing that, in the left hand side, all terms other than $(\widetilde{\lambda}_i - \lambda_j)\,\widetilde{u}_i^* u_j$ fall-off, either due to $u_i^* u_j = 0$, when $i \neq j$, or because $\delta\lambda_i = \widetilde{\lambda}_i - \lambda_j$, o.w. □

As the expression reveals, $\langle \widetilde{u}_i, u_j \rangle$ depends on the orientation of $\widetilde{u}_i$ with respect to all other u_ℓ. Moreover, the angles between eigenvectors depend not only on the minimal gap between the subspace of interest and its complement space (as in the $\sin(\Theta)$ theorem), but on every difference $\widetilde{\lambda}_i - \lambda_j$. This is a crucial ingredient to a tight bound, that will be retained throughout our analysis.

3.2 Bounding arbitrary angles

We proceed to decouple the inner products.

Theorem 3.1. *For any Hermitian matrices C and $\widetilde{C} = \delta C + C$, with eigenvectors u_j and $\widetilde{u}_i$ respectively, we have that $|\widetilde{\lambda}_i - \lambda_j|\,|\langle \widetilde{u}_i, u_j \rangle| \leq \|\delta C\, u_j\|_2$.*

Proof. We rewrite Lemma 3.1 as

$$(\widetilde{\lambda}_i - \lambda_j)^2 (\widetilde{u}_i^* u_j)^2 = \left(\sum_{\ell=1}^{n} (\widetilde{u}_i^* u_\ell)\,(u_j^* \delta C u_\ell) \right)^2 . \tag{10}$$

We now use the Cauchy-Schwartz inequality

$$\begin{aligned} (\widetilde{\lambda}_i - \lambda_j)^2 (\widetilde{u}_i^* u_j)^2 &\leq \sum_{\ell=1}^{n} (\widetilde{u}_i^* u_\ell)^2 \sum_{\ell=1}^{n} (u_j^* \delta C u_\ell)^2 \\ &= \sum_{\ell=1}^{n} (u_j^* \delta C u_\ell)^2 = \|\delta C\, u_j\|_2^2, \end{aligned} \tag{11}$$

where in the last step we exploited Lemma 3.2. The proof concludes by taking a square root at both sides of the inequality. □

Lemma 3.2. $\sum_{\ell=1}^{n} (u_j^* \delta C u_\ell)^2 = \|\delta C\, u_j\|_2^2$.

Proof. We first notice that $u_j^* \delta C u_\ell$ is a scalar and equal to its transpose. Moreover, δC is Hermitian as the difference of two Hermitian matrices. We therefore have that

$$\begin{aligned} \sum_{\ell=1}^{n} (u_j^* \delta C u_\ell)^2 &= \sum_{\ell=1}^{n} u_j^* \delta C u_\ell u_\ell^* \delta C u_j \\ &= u_j^* \delta C \sum_{\ell=1}^{n} (u_\ell u_\ell^*) \delta C u_j = u_j^* \delta C \delta C u_j = \|\delta C u_j\|_2^2, \end{aligned}$$

matching our claim. □

3.3 Refinement

As a last step, we move all perturbation terms to the numerator, at the expense of a multiplicative constant factor.

Theorem 3.2. *For Hermitian matrices C and $\widetilde{C} = \delta C + C$, with eigenvectors u_j and $\widetilde{u}_i$ respectively, the inequality*

$$|\langle \widetilde{u}_i, u_j \rangle| \leq \frac{2\,\|\delta C u_j\|_2}{|\lambda_i - \lambda_j|},$$

holds for sgn$(\lambda_i - \lambda_j)\,2\widetilde{\lambda}_i >$ sgn$(\lambda_i - \lambda_j)(\lambda_i + \lambda_j)$ and $\lambda_i \neq \lambda_j$.

Proof. Adding and subtracting λ_i from the left side of the expression in Lemma 3.1 and from definition we have

$$(\delta\lambda_i + \lambda_i - \lambda_j)\,(\widetilde{u}_i^* u_j) = \sum_{\ell=1}^{n} (\widetilde{u}_i^* u_\ell)\,(u_j^* \delta C u_\ell). \tag{12}$$

For $\lambda_i \neq \lambda_j$, the above expression can be re-written as

$$\begin{aligned} |\widetilde{u}_i^* u_j| &= \frac{\left| \sum_{\ell=1}^{n} (\widetilde{u}_i^* u_\ell)\,(u_j^* \delta C u_\ell) - \delta\lambda_i\,(\widetilde{u}_i^* u_j) \right|}{|\lambda_i - \lambda_j|} \\ &\leq 2 \max \left\{ \frac{\left| \sum_{\ell=1}^{n} (\widetilde{u}_i^* u_\ell)\,(u_j^* \delta C u_\ell) \right|}{|\lambda_i - \lambda_j|}, \frac{|\delta\lambda_i|\;|\widetilde{u}_i^* u_j|}{|\lambda_i - \lambda_j|} \right\} . \end{aligned} \tag{13}$$

Let us examine the right-hand side inequality carefully. Obviously, when the condition $|\lambda_i - \lambda_j| \le 2\,|\delta\lambda_i|$ is not met, the right clause of (13) is irrelevant. Therefore, for $|\delta\lambda_i| < |\lambda_i - \lambda_j|\,/2$ the bound simplifies to

$$|\widetilde{u}_i^* u_j| \le \frac{2\left|\sum\limits_{\ell=1}^{n} (\widetilde{u}_i^* u_\ell)\,(u_j^* \delta C u_\ell)\right|}{|\lambda_i - \lambda_j|}. \tag{14}$$

Similar to the proof of Theorem 3.1, applying the Cauchy-Schwartz inequality we have that

$$|\widetilde{u}_i^* u_j| \le \frac{2\sqrt{\sum\limits_{\ell=1}^{n} (\widetilde{u}_i^* u_\ell)^2 \sum\limits_{\ell=1}^{n} (u_j^* \delta C u_\ell)^2}}{|\lambda_i - \lambda_j|} = \frac{2\,\|\delta C u_j\|_2}{|\lambda_i - \lambda_j|}, \tag{15}$$

where in the last step we used Lemma 3.2. To finish the proof we notice that, due to Theorem 3.2, whenever $|\lambda_i - \lambda_j| \le |\widetilde{\lambda}_i - \lambda_j|$, one has

$$|\widetilde{u}_i^* u_j| \le \frac{\|\delta C\, u_j\|_2}{|\widetilde{\lambda}_i - \lambda_j|} \le \frac{\|\delta C\, u_j\|_2}{|\lambda_i - \lambda_j|} < \frac{2\,\|\delta C u_j\|_2}{|\lambda_i - \lambda_j|}. \tag{16}$$

Our bound therefore holds for the union of intervals $|\delta\lambda_i| < |\lambda_i - \lambda_j|\,/2$ and $|\lambda_i - \lambda_j| \le |\widetilde{\lambda}_i - \lambda_j|$, i.e., for $\widetilde{\lambda}_i > (\lambda_i + \lambda_j)/2$ when $\lambda_i > \lambda_j$ and for $\widetilde{\lambda}_i < (\lambda_i + \lambda_j)/2$ when $\lambda_i < \lambda_j$. □

4 Concentration of Measure

This section builds on the perturbation results of Section 3 to characterize how far any inner product $\langle \widetilde{u}_i, u_j \rangle$ and eigenvalue $\widetilde{\lambda}_i$ are from the ideal estimates.

Before proceeding, we remark on some simplifications employed in the following. W.l.o.g., we will assume that the mean $\mathbf{E}[x]$ is zero. In addition, we will assume the perspective of Theorem 3.2, for which the inequality $\text{sgn}(\lambda_i - \lambda_j)\, 2\widetilde{\lambda}_i > \text{sgn}(\lambda_i - \lambda_j)(\lambda_i + \lambda_j)$ holds. This event is shown to occur a.a.s. when the gap and the sample size are sufficiently large, but it is convenient to assume that it happens almost surely. In fact, removing this assumption is possible (see Section 4.1.2), but it is largely not pursued here as it leads to less elegant and sharp estimates.

4.1 Distributions with finite second moment

Our first flavor of results is based on a variant of the Tchebichef inequality and holds for any distribution with finite second moment, though only with moderate probability estimates.

4.1.1 Concentration of eigenvector angles

We start with the concentration of inner-products $|\langle \widetilde{u}_i, u_j \rangle|$.

Theorem 4.1. *For any two eigenvectors $\widetilde{u}_i$ and u_j of the sample and actual covariance respectively, with $\lambda_i \ne \lambda_j$, and for any real number $t > 0$, we have*

$$\mathbf{P}(|\langle \widetilde{u}_i, u_j \rangle| \ge t) \le \frac{1}{m}\left(\frac{2\,k_j}{t\,|\lambda_i - \lambda_j|}\right)^2$$

for sgn$(\lambda_i - \lambda_j)\, 2\widetilde{\lambda}_i >$ sgn$(\lambda_i - \lambda_j)(\lambda_i + \lambda_j)$ and $k_j = \left(\mathbf{E}\left[\|xx^* u_j\|_2^2\right] - \lambda_j^2\right)^{1/2}$.

Proof. According to a variant of Tchebichef's inequality (Sarwate, 2013), for any random variable X and for any real numbers $t > 0$ and α:

$$\mathbf{P}\left(|X - \alpha| \ge t\right) \le \frac{\mathbf{Var}[X] + (\mathbf{E}[X] - \alpha)^2}{t^2}. \tag{17}$$

Setting $X = \langle \widetilde{u}_i, u_j \rangle$ and $\alpha = 0$, we have

$$\begin{aligned}\mathbf{P}(|\langle \widetilde{u}_i, u_j \rangle| \ge t) &\le \frac{\mathbf{Var}[\langle \widetilde{u}_i, u_j \rangle] + \mathbf{E}[\langle \widetilde{u}_i, u_j \rangle]^2}{t^2} \\ &= \frac{\mathbf{E}\left[\langle \widetilde{u}_i, u_j \rangle^2\right]}{t^2} \le \frac{4\,\mathbf{E}\left[\|\delta C u_j\|_2^2\right]}{t^2(\lambda_i - \lambda_j)^2},\end{aligned} \tag{18}$$

where the last inequality follows from Theorem 3.2. We continue by expanding δC using the definition of the eigenvalue decomposition and substituting the expectation.

$$\begin{aligned}\mathbf{E}\left[\|\delta C u_j\|_2^2\right] &= \mathbf{E}\left[\|\widetilde{C} u_j - \lambda_j u_j\|_2^2\right] \\ &= \mathbf{E}\left[u_j^* (\widetilde{C} - \lambda_j)(\widetilde{C} - \lambda_j) u_j\right] \\ &= \mathbf{E}\left[u_j^* \widetilde{C}^2 u_j\right] + \lambda_j^2 - 2\lambda_j u_j^* \mathbf{E}\left[\widetilde{C}\right] u_j \\ &= \mathbf{E}\left[u_j^* \widetilde{C}^2 u_j\right] - \lambda_j^2.\end{aligned} \tag{19}$$

In addition,

$$\begin{aligned}\mathbf{E}\left[u_j^* \widetilde{C}^2 u_j\right] &= \sum_{p,q=1}^{m} u_j^* \frac{\mathbf{E}\left[(x_p x_p^*)(x_q x_q^*)\right]}{m^2} u_j \\ &= \sum_{p \ne q} u_j^* \frac{\mathbf{E}\left[x_p x_p^*\right] \mathbf{E}\left[x_q x_q^*\right]}{m^2} u_j + \sum_{p=1}^{m} u_j^* \frac{\mathbf{E}\left[x_p x_p^* x_p x_p^*\right]}{m^2} u_j \\ &= \frac{m(m-1)}{m^2} \lambda_j^2 + \frac{1}{m} u_j^* \mathbf{E}[xx^* xx^*]\, u_j \\ &= (1 - \frac{1}{m})\, \lambda_j^2 + \frac{1}{m} u_j^* \mathbf{E}[xx^* xx^*]\, u_j\end{aligned} \tag{20}$$

and therefore

$$\begin{aligned}\mathbf{E}\left[\|\delta C u_j\|_2^2\right] &= (1 - \frac{1}{m})\, \lambda_j^2 + \frac{1}{m} u_j^* \mathbf{E}[xx^* xx^*]\, u_j - \lambda_j^2 \\ &= \frac{u_j^* \mathbf{E}[xx^* xx^*]\, u_j - \lambda_j^2}{m} = \frac{\mathbf{E}\left[\|xx^* u_j\|_2^2\right] - \lambda_j^2}{m}.\end{aligned}$$

Putting everything together, the claim follows. □

The following corollary will be very useful when applying our results.

Corollary 4.1. *For any weights w_{ij} and real $t > 0$:*

$$\mathbf{P}\left(\sum_{i\neq j} w_{ij}\langle\widetilde{u}_i, u_j\rangle^2 > t\right) \leq \sum_{i\neq j}\frac{4\, w_{ij}\, k_j^2}{m\, t\,(\lambda_i - \lambda_j)^2},$$

where $k_j = \left(\mathbf{E}\left[\|xx^*u_j\|_2^2\right] - \lambda_j^2\right)^{1/2}$ and $w_{ij} \neq 0$ when $\lambda_i \neq \lambda_j$ and $sgn(\lambda_i - \lambda_j)\, 2\widetilde{\lambda}_i > sgn(\lambda_i - \lambda_j)(\lambda_i + \lambda_j)$.

Proof. We proceed as in the proof of Theorem 4.1:

$$\mathbf{P}\left(\Big(\sum_{i\neq j} w_{ij}\langle\widetilde{u}_i, u_j\rangle^2\Big)^{\frac{1}{2}} > t\right) \leq \frac{\mathbf{E}\left[\sum_{i\neq j} w_{ij}\langle\widetilde{u}_i, u_j\rangle^2\right]}{t^2}$$

$$\leq \frac{4}{t^2}\sum_{i\neq j} w_{ij}\frac{\mathbf{E}\left[\|\delta C u_j\|_2^2\right]}{(\lambda_i - \lambda_j)^2} \qquad (21)$$

The claim follows by computing $\mathbf{E}\left[\|\delta C u_j\|_2^2\right]$ (as before) and squaring both terms within the probability. □

4.1.2 Eigenvalue concentration

Though perhaps less sharp than what is currently known (e.g., see (Silverstein & Bai, 1995; Bai & Silverstein, 1998) for the asymptotic setting), it might be interesting to observe that a slight modification of the same argument can be used to characterize the eigenvalue relative difference, and as a consequence the main condition of Theorem 4.1.

Corollary 4.2. *For any eigenvalues λ_i and $\widetilde{\lambda}_i$ of C and $\widetilde{C}$, respectively, and for any $t > 0$, we have*

$$\mathbf{P}\left(\frac{|\widetilde{\lambda}_i - \lambda_i|}{\lambda_i} \geq t\right) \leq \frac{1}{m}\left(\frac{k_i}{\lambda_i\, t}\right)^2,$$

where $k_i = (\mathbf{E}\left[\|xx^*u_i\|_2^2\right] - \lambda_i)^{1/2}$.

Proof. Directly from the Bauer-Fike theorem (Bauer & Fike, 1960) one sees that

$$|\delta\lambda_i| \leq \|\widetilde{C}u_i - \lambda_i u_i\|_2 = \|\delta C u_i\|_2. \qquad (22)$$

The proof is then identical to that of Theorem 4.1. □

Using this, we find that the event $E = \{\text{sgn}(\lambda_i - \lambda_j)\, 2\widetilde{\lambda}_i > \text{sgn}(\lambda_i - \lambda_j)(\lambda_i + \lambda_j)\}$ occurs with probability at least

$$\mathbf{P}(E) \geq \mathbf{P}\left(|\widetilde{\lambda}_i - \lambda_i| < \frac{|\lambda_i - \lambda_j|}{2}\right) > 1 - \frac{2k_i^2}{m|\lambda_i - \lambda_j|}.$$

Therefore, one eliminates the condition from Theorem 4.1's statement by relaxing the bound to

$$\mathbf{P}(|\langle\widetilde{u}_i, u_j\rangle| \geq t) \leq \mathbf{P}(|\langle\widetilde{u}_i, u_j\rangle| \geq t \,|\, E) + (1 - \mathbf{P}(E))$$

$$< \frac{2}{m|\lambda_i - \lambda_j|}\Big(\frac{2k_j^2}{t^2|\lambda_i - \lambda_j|} + k_i^2\Big). \qquad (23)$$

4.1.3 The influence of the distribution

As seen by the straightforward inequality $\mathbf{E}\left[\|xx^*u_j\|_2^2\right] \leq \mathbf{E}\left[\|x\|_2^4\right]$, k_j connects to the kurtosis of the distribution. However, it also captures the tendency of the distribution to fall in the span of u_j.

To see this, we will work with the whitened random vectors $\varepsilon = C^{+1/2}x$, where C^+ denotes the Moore–Penrose pseudoinverse of C. In particular,

$$\begin{aligned} k_j^2 &= \mathbf{E}\left[u_j^* C^{1/2}\varepsilon\varepsilon^* C\varepsilon\varepsilon^* C^{1/2}u_j\right] - \lambda_j^2 \\ &= \lambda_j(\mathbf{E}\left[\|\Lambda^{1/2}U^*\varepsilon\varepsilon^*u_j\|_2^2\right] - \lambda_j) \\ &= \lambda_j\Big(\sum_{\ell=1}^{n}\lambda_\ell\mathbf{E}\left[\hat{\varepsilon}(\ell)^2\hat{\varepsilon}(j)^2\right] - \lambda_j\Big), \end{aligned} \qquad (24)$$

where $\hat{\varepsilon} = U^*\varepsilon$. It is therefore easier to untangle the spaces spanned by $\widetilde{u}_i$ and u_j when the variance of the distribution along the latter space is small (the expression is trivially minimized when $\lambda_j \to 0$) or when the variance is entirely contained along that space (the expression is also small when $\lambda_i = 0$ for all $i \neq j$). In addition, it can be seen that distributions with fast decaying tails allow for better principal component identification ($\mathbf{E}\left[\hat{\varepsilon}(j)^4\right]$ is a measure of kurtosis over the direction of u_j).

For the particular case of a Normal distribution, we provide a closed-form expression.

Corollary 4.3. *For a Normal distribution, we have $k_j^2 = \lambda_j\,(\lambda_j + tr(C))$.*

Proof. For a centered and normal distribution with identity covariance, the choice of basis is arbitrary and the vector $\hat{\varepsilon} = U^*\varepsilon$ is also zero mean with identity covariance. Moreover, for every $\ell \neq j$ we can write $\mathbf{E}\left[\hat{\varepsilon}(\ell)^2\hat{\varepsilon}(j)^2\right] = \mathbf{E}\left[\hat{\varepsilon}(\ell)^2\right]\mathbf{E}\left[\hat{\varepsilon}(j)^2\right] = 1$. This implies that

$$\mathbf{E}\left[\|xx^*u_j\|_2^2\right] = \lambda_j^2\,\mathbf{E}\left[\hat{\varepsilon}(j)^4\right] + \lambda_j\sum_{\ell\neq j}^{n}\lambda_\ell$$

$$= \lambda_j^2(3-1) + \lambda_j\text{tr}(C) = 2\lambda_j^2 + \lambda_j\text{tr}(C) \qquad (25)$$

and, accordingly, $k_j^2 = \lambda_j\,(\lambda_j + \text{tr}(C))$. □

4.2 Distributions supported in a Euclidean ball

Our last result provides a sharper probability estimate for the family of sub-gaussian distributions supported in a centered Euclidean ball of radius r, with their Ψ_2-norm

$$\|x\|_{\Psi_2} = \sup_{y\in\mathcal{S}^{n-1}}\|\langle x, y\rangle\|_{\psi_2}\,, \qquad (26)$$

where $\mathcal{S}^{n-1}$ is the unit sphere and with the ψ_2-norm of a random variable X defined as

$$\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}\mathbf{E}[|X|^p]^{1/p}. \qquad (27)$$

Our setting is therefore similar to the one used to study covariance estimation (Vershynin, 2012). Due to space constraints, we refer the reader to the excellent review article (Vershynin, 2010) for an introduction to sub-gaussian distributions as a tool for non-asymptotic analysis of random matrices.

Theorem 4.2. *For sub-gaussian distributions supported within a centered Euclidean ball of radius r, there exists an absolute constant c, independent of the sample size, such that for any real number $t > 0$,*

$$\mathbf{P}\left(|\langle \widetilde{u}_i, u_j\rangle| \geq t\right) \leq \mathit{exp}\left(1 - \frac{c\, m\, \Phi_{ij}(t)^2}{\lambda_j\, \|x\|_{\Psi_2}^2}\right), \tag{28}$$

where $\Phi_{ij}(t) = \frac{|\lambda_i - \lambda_j|\, t - 2\lambda_j}{2\,(r^2/\lambda_j - 1)^{1/2}} - 2\,\|x\|_{\Psi_2}$, $\lambda_i \neq \lambda_j$ and $\mathit{sgn}(\lambda_i - \lambda_j)\, 2\widetilde{\lambda}_i > \mathit{sgn}(\lambda_i - \lambda_j)(\lambda_i + \lambda_j)$.

Proof. We start from the simple observation that, for every upper bound B of $|\langle \widetilde{u}_i, u_j\rangle|$ the relation $\mathbf{P}(|\langle \widetilde{u}_i, u_j\rangle| > t) \leq \mathbf{P}(B > t)$ holds. To proceed therefore we will construct a bound with a known tail. As we saw in Sections 3.3 and 4.1,

$$\begin{aligned}
|\langle \widetilde{u}_i, u_j\rangle| &\leq \frac{2\,\|\delta C u_j\|_2}{|\lambda_i - \lambda_j|} \\
&= \frac{2\,\left\|(1/m)\sum_{p=1}^m (x_p x_p^* u_j - \lambda_j u_j)\right\|_2}{|\lambda_i - \lambda_j|} \\
&\leq \frac{2\sum_{p=1}^m \left\|x_p x_p^* u_j - \lambda_j u_j\right\|_2}{m\,|\lambda_i - \lambda_j|} \\
&= \frac{2\sum_{p=1}^m \sqrt{(u_j^* x_p)^2 (x_p^* x_p) - 2\lambda_j (u_j^* x_p)^2 + \lambda_j^2}}{m\,|\lambda_i - \lambda_j|} \\
&= \frac{2\sum_{p=1}^m \sqrt{(u_j^* x_p)^2 (\|x_p\|_2^2 - \lambda_j) + \lambda_j^2}}{m\,|\lambda_i - \lambda_j|}
\end{aligned} \tag{29}$$

Assuming further that $\|x\|_2 \leq r$, and since the numerator is minimized when $\|x_p\|_2^2$ approaches λ_j, we can write for every sample $x = C^{1/2}\varepsilon$:

$$\begin{aligned}
\sqrt{(u_j^* x)^2 (\|x\|_2^2 - \lambda_j) + \lambda_j^2} &\leq \sqrt{(u_j^* x)^2 (r^2 - \lambda_j) + \lambda_j^2} \\
&= \sqrt{\lambda_j (u_j^* \varepsilon)^2 (r^2 - \lambda_j) + \lambda_j^2} \\
&\leq |u_j^* \varepsilon|\, \sqrt{\lambda_j r^2 - \lambda_j^2} + \lambda_j,
\end{aligned} \tag{30}$$

which is a shifted and scaled version of the random variable $|\hat{\varepsilon}(j)| = |u_j^* \varepsilon|$. Setting $a = (\lambda_j r^2 - \lambda_j^2)^{1/2}$, we have

$$\begin{aligned}
\mathbf{P}\left(|\langle \widetilde{u}_i, u_j\rangle| \geq t\right) &\leq \mathbf{P}\left(\frac{2\sum_{p=1}^m (|\hat{\varepsilon}_p(j)|\, a + \lambda_j)}{m\,|\lambda_i - \lambda_j|} \geq t\right) \\
&= \mathbf{P}\left(\sum_{p=1}^m (|\hat{\varepsilon}_p(j)|\, a + \lambda_j) \geq 0.5\, mt\, |\lambda_i - \lambda_j|\right) \\
&= \mathbf{P}\left(\sum_{p=1}^m |\hat{\varepsilon}_p(j)| \geq \frac{m\,(0.5\, t\, |\lambda_i - \lambda_j| - \lambda_j)}{a}\right).
\end{aligned} \tag{31}$$

By Lemma 4.1 however, the left hand side is a sum of independent sub-gaussian variables. Since the summands are not centered, we expand each $|\hat{\varepsilon}_p(j)| = z_p + \mathbf{E}[|\hat{\varepsilon}_p(j)|]$ in terms of a centered sub-gaussian z_p with the same ψ_2-norm. Furthermore, by Jensen's inequality and Lemma 4.1

$$\mathbf{E}[|\hat{\varepsilon}_p(j)|] \leq \mathbf{E}\left[\hat{\varepsilon}_p(j)^2\right]^{1/2} \leq \frac{2}{\lambda_j}\, \|x\|_{\Psi_2}. \tag{32}$$

Therefore, if we set $\Phi_{ij}(t) = \frac{(0.5\,|\lambda_i - \lambda_j|\, t - \lambda_j)}{(r^2/\lambda_j - 1)^{1/2}} - 2\,\|x\|_{\Psi_2}$

$$\mathbf{P}\left(|\langle \widetilde{u}_i, u_j\rangle| \geq t\right) \leq \mathbf{P}\left(\sum_{p=1}^m z_p \geq \frac{m\Phi_{ij}(t)}{\lambda_j}\right). \tag{33}$$

Moreover, by the rotation invariance principle, the left hand side of the last inequality is a sub-gaussian with ψ_2-norm smaller than $(c_1 \sum_{p=1}^m \|z_p\|_{\psi_2}^2)^{1/2} = (c_1 m)^{1/2}\, \|z\|_{\psi_2} \leq (c_1 m/\lambda_j)^{1/2}\, \|x\|_{\Psi_2}$, for some absolute constant c_1. As a consequence, there exists an absolute constant c_2, such that for each $\theta > 0$:

$$\mathbf{P}\left(\left|\sum_{p=1}^m z_p\right| \geq \theta\right) \leq \exp\left(1 - \frac{c_2\, \theta^2 \lambda_j}{m\, \|x\|_{\Psi_2}^2}\right). \tag{34}$$

Substituting $\theta = m\, \Phi_{ij}(t)/\lambda_j$, we have

$$\begin{aligned}
\mathbf{P}\left(|\langle \widetilde{u}_i, u_j\rangle| \geq t\right) &\leq \exp\left(1 - \frac{c_2\, m^2\, \Phi_{ij}(t)^2 \lambda_j}{m \lambda_j^2\, \|x\|_{\Psi_2}^2}\right) \\
&= \exp\left(1 - \frac{c_2\, m\, \Phi_{ij}(t)^2}{\lambda_j\, \|x\|_{\Psi_2}^2}\right),
\end{aligned} \tag{35}$$

which is the desired bound. □

Lemma 4.1. *If x is a sub-gaussian random vector and $\varepsilon = C^{+1/2}x$, then for every i, the random variable $\hat{\varepsilon}(i) = u_i^* \varepsilon$ is also sub-gaussian, with $\|\hat{\varepsilon}(i)\|_{\psi_2} \leq \|x\|_{\Psi_2} / \sqrt{\lambda_i}$.*

Proof. Notice that

$$\begin{aligned}
\|x\|_{\Psi_2} = \sup_{y \in \mathcal{S}^{n-1}} \|\langle x, y\rangle\|_{\psi_2} &= \sup_{y \in \mathcal{S}^{n-1}} \left\|\sum_{j=1}^n \lambda_j^{1/2} (u_j^* y)(u_j^* \varepsilon)\right\|_{\psi_2} \\
&\geq \left\|\sum_{j=1}^n \lambda_j^{1/2} (u_j^* u_i) \hat{\varepsilon}(j)\right\|_{\psi_2} = \lambda_i^{1/2}\, \|\hat{\varepsilon}(i)\|_{\psi_2},
\end{aligned} \tag{36}$$

where, for the last inequality, we set $y = u_i$. □

5 Application to Dimensionality Reduction

To emphasize the utility of our results, in the following we consider the practical example of linear dimensionality reduction. We show that a direct application of our bounds leads to upper estimates on the sample requirement.

In terms of mean squared error, the optimal way to reduce the dimension of a sample x of a distribution is by projecting it over the subspace of the covariance with maximum variance. Denote by I_k the diagonal matrix with the first k diagonal entries equal to one and the rest zero. When the actual covariance is known, the expected energy loss induced by the $P_k x = I_k U^* x$ projection is

$$\mathit{loss}(P_k) = \frac{\mathbf{E}\left[\|x\|_2^2 - \|P_k x\|_2^2\right]}{\mathbf{E}\left[\|x\|_2^2\right]} = \frac{\sum_{i>k} \lambda_i}{\mathrm{tr}(C)}. \quad (37)$$

However, when the projector $\widetilde{P}_k = I_k \widetilde{U}^*$ is constructed from the sample covariance, we have

$$\begin{aligned} \mathit{loss}(\widetilde{P}_k) &= \frac{\mathbf{E}\left[\|x\|_2^2 - \|\widetilde{P}_k x\|_2^2\right]}{\mathbf{E}\left[\|x\|_2^2\right]} \\ &= \frac{\sum_{i=1}^n \lambda_i - \mathrm{tr}(I_k \widetilde{U}^* U \Lambda U^* \widetilde{U})}{\mathrm{tr}(C)} \\ &= \frac{\sum_{i=1}^n \lambda_i - \sum_{i\leq k, j} (\widetilde{u}_i^* u_j)^2 \lambda_j}{\mathrm{tr}(C)} \end{aligned} \quad (38)$$

with the expectation taken over the to-be-projected vectors x, but not the samples used to estimate the covariance. After slight manipulation, one finds that

$$\mathit{loss}(\widetilde{P}_k) = \mathit{loss}(P_k) + \frac{\sum\limits_{i\leq k, j\neq i} (\widetilde{u}_i^* u_j)^2 (\lambda_i - \lambda_j)}{\mathrm{tr}(C)}. \quad (39)$$

The loss difference has an intuitive interpretation: when reducing the dimension with $\widetilde{P}_k$ one looses either by discarding useful energy (terms $j > k$), or by displacing kept components within the permissible eigenspace (terms $j \leq k$). Note also that all terms with $j < i$ are negative and can be excluded from the sum if we are satisfied we an upper estimate[2].

It is an implication of (39) and Corollary 4.1 that, when its conditions hold, for any distribution and $t > 0$

$$\mathbf{P}\left(\mathit{loss}(\widetilde{P}_k) > \mathit{loss}(P_k) + \frac{t}{\mathrm{tr}(C)} \right) \leq \sum_{\substack{i\leq k \\ j>i}} \frac{4\, k_j^2}{mt\, |\lambda_i - \lambda_j|}.$$

Observe that the loss difference becomes particularly small whenever k is small: (i) the terms in the sum are fewer and (ii) the magnitude of each term decreases (due to $|\lambda_i - \lambda_j|$).

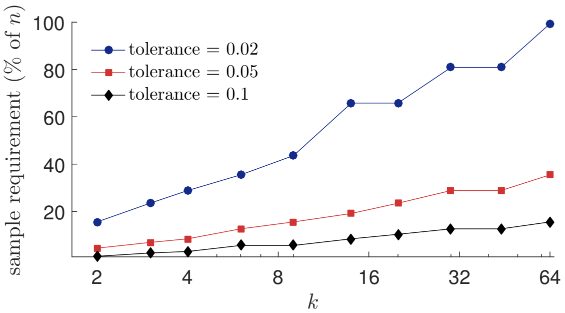


Figure 2: The figure depicts for each k, the sample size needed such that the loss difference $\mathit{loss}(\widetilde{P}_k) - \mathit{loss}(P_k)$ becomes smaller than some tolerance. We can observe that, in MNIST, linear dimensionality reduction works with fewer than $n = 725$ samples when the size k of the reduced dimension is small.

This phenomenon is also numerically verified in Figure 2 for the distribution of the images featuring digit '3' in MNIST (total 6131 images with $n = 784$ pixels each). The figure depicts for different k how many samples are required such that the loss difference is smaller than a tolerance threshold, here 0.02, 0.05, and 0.1. Each point in the figure corresponds to an average over 10 sampling draws. The trends featured in these numerical results agree with our theoretical intuition. Moreover they illustrate that for modest k the sample requirement is far smaller than n.

It is also interesting to observe that for covariance matrices that are (approximately) low-rank, we obtain estimates reminiscent of compressed sensing (Candès et al., 2011), in the sense that the sample requirement becomes a function of the non-zero eigenvalues. Though intuitive, with the exception of (Koltchinskii et al., 2016), this dependency of the estimation accuracy on the rank was not transparent in known results for covariance estimation (Rudelson, 1999; Adamczak et al., 2010; Vershynin, 2012).

6 Conclusions

The main contribution of this paper was the derivation of non-asymptotic bounds for the concentration of inner-products $|\langle \widetilde{u}_i, u_j \rangle|$ involving eigenvectors of the sample and actual covariance matrices. We also showed how these results can be extended to reason about eigenvalues and we applied them to the non-asymptotic analysis of linear dimensionality reduction.

We have identified two interesting directions for further research. The first has to do with obtaining tighter estimates. Especially with regards to our perturbation arguments, we believe that our current bounds on inner products could be sharpened by at least a constant multiplicative factor. The second direction involves using our results for the analysis of methods that utilize the eigenvectors of the covariance, such that principal component projection and regression (Jolliffe, 1982; Frostig et al., 2016).

[2] A similar approach could also be utilized to derive a lower bound of the quantity $\mathit{loss}(\widetilde{P}_k) - \mathit{loss}(P_k)$.

References

Adamczak, Radosław, Litvak, Alexander, Pajor, Alain, and Tomczak-Jaegermann, Nicole. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561, 2010.

Ahmed, SE. Large-sample estimation strategies for eigenvalues of a wishart matrix. *Metrika*, 47(1):35–45, 1998.

Anderson, Theodore Wilbur. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148, 1963.

Bai, ZD. Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica*, pp. 611–662, 1999.

Bai, ZD and Yin, YQ. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The annals of Probability*, pp. 1275–1294, 1993.

Bai, ZD, Miao, BQ, Pan, GM, et al. On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability*, 35(4):1532–1572, 2007.

Bai, Zhi-Dong and Silverstein, Jack W. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Annals of probability*, pp. 316–345, 1998.

Bauer, Friedrich L and Fike, Charles T. Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141, 1960.

Berkmann, Jens and Caelli, Terry. Computation of surface geometry and segmentation using covariance techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(11):1114–1116, 1994.

Candès, Emmanuel J., Li, Xiaodong, Ma, Yi, and Wright, John. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, June 2011.

Davis, Chandler and Kahan, William Morton. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Frostig, Roy, Musco, Cameron, Musco, Christopher, and Sidford, Aaron. Principal component projection without principal component analysis. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 2349–2357, 2016.

Girko, V. Strong law for the eigenvalues and eigenvectors of empirical covariance matrices. 1996.

Huang, Ling, Yan, Donghui, Taft, Nina, and Jordan, Michael I. Spectral clustering with perturbed data. In *Advances in Neural Information Processing Systems*, pp. 705–712, 2009.

Hunter, Blake and Strohmer, Thomas. Performance analysis of spectral clustering on compressed, incomplete and inaccurate measurements. *arXiv preprint arXiv:1011.0997*, 2010.

Jolliffe, Ian. *Principal component analysis*. Wiley Online Library, 2002.

Jolliffe, Ian T. A note on the use of principal components in regression. *Applied Statistics*, pp. 300–303, 1982.

Kambhatla, Nandakishore and Leen, Todd K. Dimension reduction by local principal component analysis. *Neural computation*, 9(7):1493–1516, 1997.

Koltchinskii, Vladimir and Lounici, Karim. Normal approximation and concentration of spectral projectors of sample covariance. *arXiv preprint arXiv:1504.07333*, 2015.

Koltchinskii, Vladimir, Lounici, Karim, et al. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pp. 1976–2013. Institut Henri Poincaré, 2016.

Mestre, Xavier. Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Transactions on Information Theory*, 54(11), 2008.

Rudelson, Mark. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.

Sarwate, Dilip. Two-sided chebyshev inequality for event not symmetric around the mean? Mathematics Stack Exchange, 2013. URL:http://math.stackexchange.com/q/144675 (version: 2012-05-13).

Schott, James R. Asymptotics of eigenprojections of correlation matrices with some applications in principal components analysis. *Biometrika*, pp. 327–337, 1997.

Shaghaghi, Mahdi and Vorobyov, Sergiy A. Subspace leakage analysis of sample data covariance matrix. In *ICASSP*, pp. 3447–3451. IEEE, 2015.

Silverstein, Jack W and Bai, ZD. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54 (2):175–192, 1995.

Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2010.

Vershynin, Roman. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Yu, Yi, Wang, Tengyao, Samworth, Richard J, et al. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.