CHAPTER 8

Principal component analysis in high dimensions

Principal component analysis (PCA) is a standard technique for exploratory data analysis and dimension reduction. It is based on seeking the maximal variance components of a distribution, or equivalently, a low-dimensional subspace that captures the majority of the variance. Given a finite collection of samples, the empirical form of principal component analysis involves computing some subset of the top eigenvectors of the sample covariance matrix. Of interest is when these eigenvectors provide a good approximation to to the subspace spanned by the top eigenvectors of the population covariance matrix. In this chapter, we study these issues in a high-dimensional and non-asymptotic framework, both for classical unstructured forms of PCA as well as more modern structured variants.

■ 8.1 Principal components and dimension reduction

Let X be a d-dimensional random vector, say with a zero mean vector and covariance matrix $\mathbf{\Sigma}$. We use $\gamma_1(\mathbf{\Sigma}) \geq \gamma_2(\mathbf{\Sigma}) \geq \cdots \geq \gamma_d(\mathbf{\Sigma}) \geq 0$ to denote the ordered eigenvalues of the covariance matrix. In its simplest instantiation, principal components analysis asks: along what unit-norm vector $v \in \mathbb{S}^{d-1}$ is the variance of the random variable $\langle v, X \rangle$ maximized? This direction is known as the first principal component at the population level, assumed here for the sake of discussion to be unique. In analytical terms, we have

$$v^* = \arg \max_{v \in \mathbb{S}^{d-1}} \operatorname{var}(\langle v, X \rangle) \; = \; \arg \max_{v \in \mathbb{S}^{d-1}} \mathbb{E}\big[\langle v, X \rangle^2\big] \; = \; \arg \max_{v \in \mathbb{S}^{d-1}} \langle v, \mathbf{\Sigma} v \rangle, \qquad (8.1)$$

so that by definition, the first principal component is the maximum eigenvector of the covariance matrix $\mathbf{\Sigma}$. More generally, we can define the top r principal components at the population level by seeking an orthonormal matrix $\mathbf{V} \in \mathbb{R}^{d \times r}$, formed with unit-norm and orthogonal columns $\{v_1, \ldots, v_r\}$, that maximizes the quantity

$$\mathbb{E}\|\mathbf{V}^T X\|_2^2 = \sum_{j=1}^{r} \mathbb{E}\big[\langle v_j,\, X\rangle^2\big]. \tag{8.2}$$

As we explore in Exercise 8.4, these principal components are simply the top r eigenvectors of the population covariance matrix $\mathbf{\Sigma}$.

In practice, however, we do not know the covariance matrix, but rather only have access to a finite collection of samples, say $\{x_i\}_{i=1}^n$, each drawn according to $\mathbb{P}$. Based on these samples (and using the zero mean assumption), we can form the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$. The empirical version of PCA is based on the "plug-in" principle, namely replacing the unknown population covariance $\mathbf{\Sigma}$ with this empirical version $\widehat{\mathbf{\Sigma}}$. For instance, the empirical analog of the first principal component (8.1) is given by the optimization problem

$$\widehat{v} = \arg\max_{v \in \mathbb{S}^{d-1}} \langle v,\, \widehat{\mathbf{\Sigma}} v\rangle. \tag{8.3}$$

Consequently, from the statistical point of view, we need to understand in what sense the minimizers of these empirically defined problems provide good approximations to their population analogues. Alternatively phrased, this question is equivalent to asking how the eigenstructures of the population and sample covariances are related, a question to be addressed later in this chapter.

■ 8.1.1 Interpretations and uses of PCA

Before turning to the analysis of PCA, let us consider some of its interpretations and applications.

Example 8.1 (PCA as matrix approximation)**.** Principal components analysis can be interpreted in terms of low-rank approximation. In particular, given som unitarily invariant matrix [1] norm $|\!|\!|\cdot|\!|\!|$, consider the problem of finding the best rank r approximation to a matrix $\mathbf{\Sigma}$

$$\mathbf{Z}^* = \arg\min_{\text{rank}(\mathbf{Z})=r} \Big\{ |\!|\!|\mathbf{\Sigma} - \mathbf{Z}|\!|\!|^2 \Big\}. \tag{8.4}$$

In this interpretation, the matrix $\mathbf{\Sigma}$ need only be symmetric, not necessarily positive semidefinite as it must be when it is a covariance matrix. A classical result known as the *Eckart-Young-Mirsky theorem* guarantees that an optimal solution $\mathbf{\Sigma}^*$ exists, and takes the form of a truncated eigendecomposition, specified in terms of the top r eigenvectors of the matrix $\mathbf{\Sigma}$. More precisely, recall that the symmetric matrix $\mathbf{\Sigma}$

[1] For a symmetric matrix M, a matrix norm is unitarily invariant if $|\!|\!|\mathbf{M}|\!|\!| = |\!|\!|\mathbf{V}^T\mathbf{M}\mathbf{V}|\!|\!|$ for any orthonormal matrix $\mathbf{V}$. See Exercise 8.2 for further discussion.

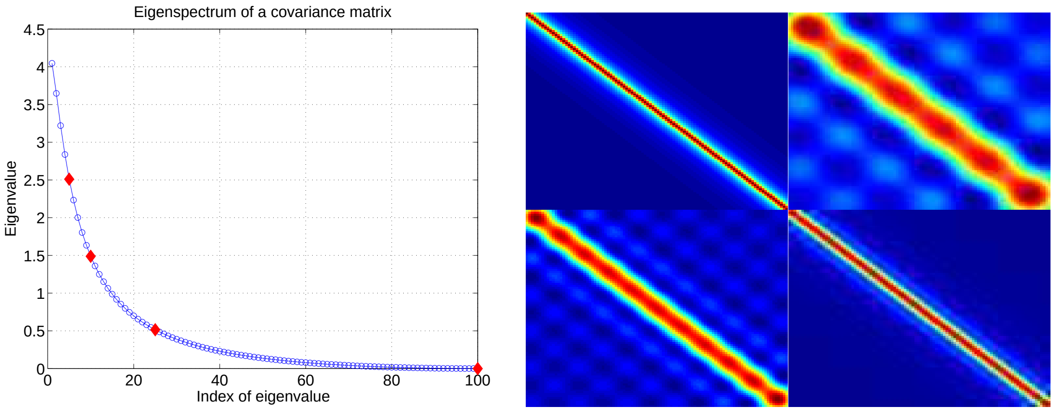
has an orthogonal basis of eigenvectors, say $\{v_1, \ldots, v_d\}$, associated with its ordered eigenvalues $\{\gamma_j(\mathbf{\Sigma})\}_{j=1}^d$. In terms of this notation, the optimal rank r approximation takes the form

$$\mathbf{\Sigma}^* = \sum_{j=1}^{r} \gamma_j(\mathbf{\Sigma}) \big(v_j \otimes v_j\big), \tag{8.5}$$

where $v_j \otimes v_j := v_j v_j^T$ is the rank one outer product. For the Frobenius matrix norm $|\!|\!|\mathbf{M}|\!|\!|_{\mathrm{F}} = \sqrt{\sum_{j,k=1}^d M_{jk}^2}$, the error in the optimal approximation is given by

$$|\!|\!|\mathbf{\Sigma}^* - \mathbf{\Sigma}|\!|\!|_{\mathrm{F}}^2 = \sum_{j=r+1}^{d} \gamma_j^2(\mathbf{\Sigma}). \tag{8.6}$$

Figure 8-1 provides an illustration of the matrix approximation view of PCA. We first generated the Toeplitz matrix $\mathbf{T} \in \mathcal{S}_+^{d \times d}$ with entries $T_{jk} = e^{-\alpha\sqrt{j-k}}$ with $\alpha = 0.95$, and then formed the recentered matrix $\mathbf{\Sigma} := \mathbf{T} - \gamma_{\min}(\mathbf{T})\mathbf{I}_d$. Panel (a) shows the eigenspectrum of the matrix $\mathbf{\Sigma}$: note that the rapid decay of the eigenvalues that renders it amenable to an accurate low-rank approximation. The top left image in panel (b) corresponds to the original matrix $\mathbf{\Sigma}$, whereas the remaining images illustrate approximations with increasing rank ($r = 5$ in top right; $r = 10$ in bottom left and



(a) (b)

Figure 8-1. Illustration of PCA for low-rank matrix approximation. (a) Eigenspectrum of a matrix $\mathbf{\Sigma} \in \mathcal{S}_+^{100 \times 100}$ generated as described in the text. Note the extremely rapid decay of the sorted eigenspectrum. Red diamonds mark the rank cutoffs $r \in \{5, 10, 25, 100\}$, which define three approximations to the whole matrix ($r = 100$.) (b) Top left: original matrix. Top right: approximation based on $r = 5$ components. Bottom left: approximation based on $r = 10$ components. Bottom right: approximation based on $r = 25$ components.

$r = 25$ in bottom right.) Although the defects in approximations with rank $r = 5$ or $r = 10$ are readily apparent, the approximation with rank $r = 25$ seems reasonable. ♣

Example 8.2 (PCA for data compression)**.** Principal component analysis can also be interpreted as a linear form of data compression. Given a zero-mean random vector $X \in \mathbb{R}^d$, a simple way in which to compress it is via projection to a lower dimensional subspace $\mathbb{S}$—say via a projection operator of the form $\Pi_{\mathbb{S}}(X)$. For a fixed dimension r, how to choose the subspace $\mathbb{S}$? Consider the criterion that chooses $\mathbb{S}$ by minimizing the mean-squared error

$$\mathbb{E}\big[\|X - \Pi_{\mathbb{S}}(X)\|_2^2\big].$$

This optimal subspace need not be unique in general, but will be when there is a gap between the eigenvalues $\gamma_r(\mathbf{\Sigma})$ and $\gamma_{r+1}(\mathbf{\Sigma})$. In this case, the optimal subspace $\mathbb{S}^*$ is spanned by the top r eigenvectors of the matrix $\mathbf{\Sigma} = \operatorname{cov}(X)$. In particular, the projection operator $\Pi_{\mathbb{S}^*}$ can be written as $\Pi_{\mathbb{S}^*}(x) = \mathbf{V}_r \mathbf{V}_r^T$, where $\mathbf{V}_r \in \mathbb{R}^{d \times r}$ is an orthogonal matrix with the top r eigenvectors $\{v_1, \ldots, v_r\}$ as its columns. Using this optimal projection $\mathbb{S}^*$, the minimal reconstruction error based on a rank r projection is given by

$$\mathbb{E}\big[\|X - \Pi_{\mathbb{S}}(X)\|_2^2\big] \;=\; \sum_{j=r+1}^{d} \gamma_j^2(\mathbf{\Sigma}), \tag{8.7}$$

where $\{\gamma_j(\mathbf{\Sigma})\}_{j=1}^d$ are the ordered eigenvalues of $\mathbf{\Sigma}$. See Exercise 8.4 for further exploration of these and other properties.

The problem of face analysis provides an interesting illustration of PCA for reconstruction or data compression. Consider a large database of face images, such as those illustrated in panel (a) of Figure 8-2. Taken from the Yale Face Database, each image is gray-scale with dimensions 243 × 320. By vectorizing each image, we obtain a vector x in $d = 243 \times 320 = 77760$ dimensions. We compute the average image $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$ based on $n = 165$ samples. Panel (b) shows the relatively fast decay of the first 100 eigenvalues of this sample covariance matrix. Panel (c) shows the average face (top left image) along with the first 24 "eigenfaces", meaning the top 25 eigenvectors of the sample covariance matrix, each converted back to a 243 × 320 image. Finally, for a particular sample, panel (d) shows a sequence of reconstructions of a given face, starting with the average fact (top left image), and followed by the average face in conjunction with principal components 1 through 24. ♣

Principal component analysis can also be used an estimator for estimation in mixture models.

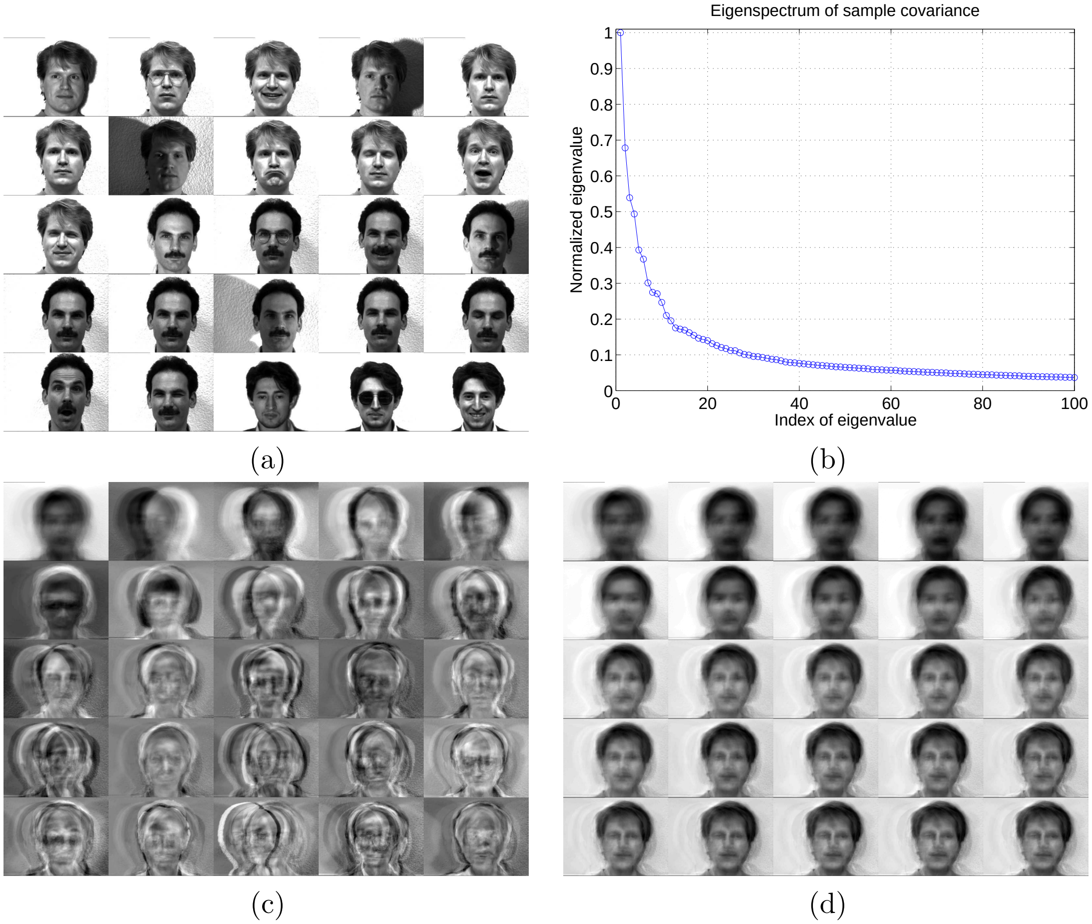
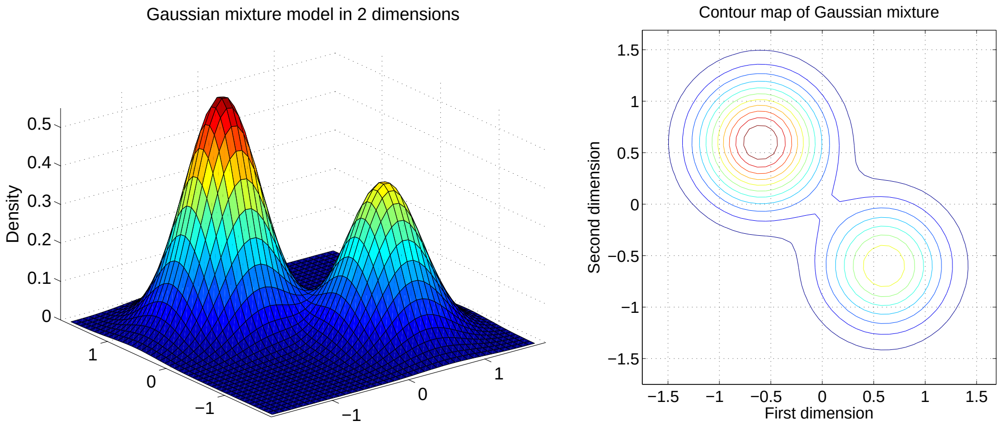


Figure 8-2. (a) Samples of face images from the Yale Face Database. (b) First 100 eigenvalues of the sample covariance matrix. (c) First 25 eigenfaces computed from the sample covariance matrix. (d) Reconstructions based on the first 25 eigenfaces plus the average face.

Example 8.3 (PCA for Gaussian mixture models)**.** Let $\phi(\cdot; \mu, \Gamma)$ denote the density of Gaussian random vector with mean μ and covariance matrix Γ. A two-component Gaussian mixture model with isotropic covariance structure is a random vector $X \in \mathbb{R}^d$ drawn according to the density

$$f(x; \theta) = \alpha \, \phi(x; -\theta^*, \sigma^2 \mathbf{I}_d) + (1 - \alpha) \, \phi(x; \theta^*, \sigma^2 \mathbf{I}_d), \tag{8.8}$$

where $\theta^* \in \mathbb{R}^d$ is a vector parameterizing the means of the two Gaussian components, $\alpha \in (0, 1)$ is a mixture weight, and $\sigma > 0$ is a dispersion term. Figure 8-3 provides an illustration of such a mixture model in $d = 2$ dimensions, with mean vector $\theta^* = \begin{bmatrix} 0.6 & -0.6 \end{bmatrix}^T$, standard deviation $\sigma = 0.4$, and weight $\alpha = 0.4$. Given samples



(a) (b)

Figure 8-3. Use of PCA for Gaussian mixture models. (a) Density function of a two-component Gaussian mixture (8.8) with mean vector $\theta^* = [0.6 \;\; -0.6]^T$, standard deviation $\sigma = 0.4$, and weight $\alpha = 0.4$. (b) Contour plots of the density function, which provide intuition as to why PCA should be useful in recovering the mean vector θ^*.

$\{x_i\}_{i=1}^n$ drawn from such a model, a natural goal is to estimate the mean vector θ^*, assuming for simplicity that the variance σ^2 is a known quantity. Principal component analysis provides a natural method for doing so. In particular, a straightforward calculation yields that

$$\mathbf{\Sigma} := \text{cov}(X) \;=\; \theta^* \otimes \theta^* + \sigma^2 \mathbf{I}_d,$$

where $\theta^* \otimes \theta^* := \theta^*(\theta^*)^T$ is a rank one outer product. Thus, we see that θ^* is proportional to the maximal eigenvector of $\mathbf{\Sigma}$. Consequently, a reasonable estimator $\widehat{\theta}$ is given by the maximal eigenvector of the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$. We study the properties of this estimator in Exercise 8.6. ♣

■ 8.1.2 Perturbations of eigenvalues and eigenspaces

Thus far, we have seen that the eigenvectors of population and sample covariance matrices are interesting objects with a range of uses. In practice, PCA is always applied to the sample covariance matrix, and the central question of interest is how well the sampled-based eigenvectors approximate those of the population covariance.

Before addressing this question, let us make a brief detour into matrix perturbation theory. Let us consider the following general question: given a symmetric matrix $\mathbf{R}$, how does its eigenstructure relate to the perturbed matrix $\mathbf{Q} = \mathbf{R} + \mathbf{P}$? Here $\mathbf{P}$ is another symmetric matrix, playing the role of the perturbation. We will see that the eigenvalues of $\mathbf{Q}$ and $\mathbf{R}$ are related in a straightforward manner. Understanding how the eigenspaces change, however, requires some more care.

Let us begin with changes in the eigenvalues. From the standard variational definition of the maximum eigenvalue, we have

$$\gamma_1(\mathbf{Q}) = \max_{v \in \mathbb{S}^{d-1}} \langle v,\, \big(\mathbf{R} + \mathbf{P}\big)v\rangle \;\leq\; \max_{v \in \mathbb{S}^{d-1}} \langle v,\, \mathbf{R}v\rangle + \max_{v \in \mathbb{S}^{d-1}} \langle v,\, \mathbf{P}v\rangle \;\leq\; \gamma_1(\mathbf{R}) + |\!|\!|\mathbf{P}|\!|\!|_{\text{op}}.$$

Since the same argument holds with the roles of $\mathbf{Q}$ and $\mathbf{R}$ reversed, we conclude that $|\gamma_1(\mathbf{Q}) - \gamma_1(\mathbf{R})| \leq |\!|\!|\mathbf{Q} - \mathbf{R}|\!|\!|_{\text{op}}$. Thus, the maximum eigenvalues of $\mathbf{Q}$ and $\mathbf{R}$ can differ by at most the operator norm of their difference. More generally, we have

$$\max_{j=1,\ldots,d} \big|\gamma_j(\mathbf{Q}) - \gamma_j(\mathbf{R})\big| \leq |\!|\!|\mathbf{Q} - \mathbf{R}|\!|\!|_{\text{op}}. \tag{8.9}$$

This bound is an instance of what is known as *Weyl's inequality*; we work through its proof in Exercise 8.3.

Although eigenvalues are generically stable, the same does not hold for eigenvectors and eigenspaces, unless further conditions are imposed. The following example provides an illustration of such instability:

Example 8.4 (Sensitivity of eigenvectors)**.** For a parameter $\epsilon \in [0, 1]$, consider the family of symmetric matrices

$$\mathbf{Q}_\epsilon := \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1.01 \end{bmatrix} \;=\; \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1.01 \end{bmatrix}}_{\mathbf{Q}_0} + \epsilon \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{\mathbf{P}}. \tag{8.10}$$

Thus, the matrix $\mathbf{Q}_\epsilon$ is a perturbation of a diagonal matrix $\mathbf{Q}_0$ by an ϵ-multiple of the fixed matrix $\mathbf{P}$. Since $|\!|\!|\mathbf{P}|\!|\!|_{\text{op}} = 1$, the magnitude of the perturbation is directly controlled by ϵ. On one hand, the eigenvalues remain stable to this perturbation: in terms of the shorthand $a = 1.01$, we have $\gamma(\mathbf{Q}_0) = \{1, a\}$ and

$$\gamma(\mathbf{Q}_\epsilon) = \Big\{\frac{1}{2}\big\{(a+1) + \sqrt{(a-1)^2 + 4\epsilon^2}\big\},\quad \frac{1}{2}\big\{(a+1) - \sqrt{(a-1)^2 + 4\epsilon^2}\big\}\Big\}.$$

Thus, we find that

$$\max_{j=1,2} \big|\gamma_j(\mathbf{Q}_0) - \gamma_j(\mathbf{Q}_\epsilon)\big| = \frac{1}{2}\Big|(a-1) - \sqrt{(a-1)^2 + 4\epsilon^2}\;\Big| \;\leq \epsilon,$$

which confirms the validity of Weyl's inequality (8.9) in this particular case.

On the other hand, the maximal eigenvector of $\mathbf{Q}_\epsilon$ is very different from that of $\mathbf{Q}_0$, even for relatively small values of ϵ. For $\epsilon = 0$, the matrix $\mathbf{Q}_0$ has the unique maximal eigenvector $v_0 = [0 \quad 1]^T$. However, if we set $\epsilon = 0.01$, a numerical calculation shows that the maximal eigenvector of $\mathbf{Q}_\epsilon$ is $v_\epsilon \approx [0.53 \quad 0.85]^T$. Note that $\|v - v_\epsilon\|_2 \gg \epsilon$, showing that eigenvectors can be extremely sensitive to perturbations. ♣

What is the underlying problem? The issue is that while $\mathbf{Q}_0$ has a unique maximal eigenvector, the gap between the largest eigenvalue $\gamma_1(\mathbf{Q}_0) = 1.01$ and its second largest $\gamma_2(\mathbf{Q}_0) = 1$ is very small. Consequently, even small perturbations of the matrix lead to "mixing" between the spaces spanned by the top and second largest eigenvectors. On the other hand, if this eigengap can be bounded away from zero, then it turns out that we can guarantee stability of the eigenvectors. We now turn to this type of theory.

■ 8.2 Bounds for generic eigenvectors

We begin our exploration of eigenvector bounds with the generic case, in which no additional structure is imposed on the eigenvectors. In later sections, we turn to structured variants of eigenvector estimation.

■ 8.2.1 A general deterministic result

Consider a symmetric positive semidefinite $\mathbf{\Sigma}$ with eigenvalues ordered as

$$\gamma_1(\mathbf{\Sigma}) \geq \gamma_2(\mathbf{\Sigma}) \geq \gamma_3(\mathbf{\Sigma}) \geq \cdots \gamma_d(\mathbf{\Sigma}) \geq 0.$$

Let $\theta^* \in \mathbb{R}^d$ denote its maximal eigenvector, assumed to be unique. Now consider a perturbed version $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma} + \mathbf{P}$ of the original matrix. As suggested by our notation, in the context of PCA, the original matrix corresponds to the population covariance matrix, whereas the perturbed matrix corresponds to the sample covariance. However, at least for the time being, our theory should be viewed as general.

As should be expected based on Example 8.4, any theory relating the maximum eigenvectors of $\mathbf{\Sigma}$ and $\widehat{\mathbf{\Sigma}}$ should involve the *eigengap* $\nu := \gamma_1(\mathbf{\Sigma}) - \gamma_2(\mathbf{\Sigma})$, assumed to be strictly positive. In addition, the following result involves the transformed perturbation matrix

$$\widetilde{\mathbf{P}} := \mathbf{U}^T \mathbf{P} \mathbf{U} \; = \; \begin{bmatrix} \tilde{p}_{11} & \tilde{p}^T \\ \tilde{p} & \widetilde{\mathbf{P}}_{22} \end{bmatrix}, \tag{8.11}$$

where $\tilde{p}_{11} \in \mathbb{R}$, $\tilde{p} \in \mathbb{R}^{d-1}$, and $\widetilde{\mathbf{P}}_{22} \in \mathbb{R}^{(d-1)\times(d-1)}$. Here $\mathbf{U}$ is an orthonormal matrix with the eigenvectors of $\mathbf{\Sigma}$ as its columns.

Theorem 8.1. Consider a positive semidefinite matrix $\mathbf{\Sigma}$ with maximum eigenvector $\theta^* \in \mathbb{S}^{d-1}$ and eigengap $\nu = \gamma_1(\mathbf{\Sigma}) - \gamma_2(\mathbf{\Sigma})$. Given a bounded perturbation $|\!|\!|\mathbf{P}|\!|\!|_{\text{op}} \leq \nu/4$, the matrix $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma} + \mathbf{P}$ has a unique (up to sign) maximal eigenvector $\widehat{\theta}$ satisfying the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{8\,\|\tilde{p}\|_2}{\nu}. \tag{8.12}$$

Proof. Our proof is variational in nature, based on the optimization problem $\max\limits_{\theta \in \mathbb{S}^{d-1}} \theta^T \widehat{\mathbf{\Sigma}} \theta$ that characterizes the maximal eigenvector of the matrix $\widehat{\mathbf{\Sigma}}$. Define the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$, and the function

$$\Psi(\Delta; \mathbf{P}) := \langle \Delta,\, \mathbf{P}\Delta \rangle + 2\langle \Delta,\, \mathbf{P}\theta^* \rangle. \tag{8.13}$$

In parallel to our analysis of sparse linear regression from Chapter 7, the first step in our analysis is to prove the *basic inequality for PCA*. For future reference, we state this inequality in a slightly more general form required for the current proof. In particular, given any subset $\mathcal{C} \subseteq \mathbb{S}^{d-1}$, let θ^* and $\widehat{\theta}$ maximize the quadratic objectives

$$\max_{\theta \in \mathcal{C}} \langle \theta,\, \mathbf{\Sigma}\theta \rangle, \quad \text{and} \quad \max_{\theta \in \mathcal{C}} \langle \theta,\, \mathbf{P}\theta \rangle, \tag{8.14}$$

respectively. The current proof involves the choice $\mathcal{C} = \mathbb{S}^{d-1}$.

Lemma 1 (PCA basic inequality)**.** *Given a matrix $\mathbf{\Sigma}$ with eigengap $\nu > 0$, the error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ is bounded as*

$$\frac{\nu}{2} \|\widehat{\Delta}\|_2^2 \leq \big|\Psi(\widehat{\Delta}; \mathbf{P})\big|. \tag{8.15}$$

Taking this inequality as given for the moment, the remainder of the proof is straightforward. By the definition (8.13) and the triangle inequality, we have

$$|\Psi(\widehat{\Delta}; \mathbf{P})| \leq |\!|\!|\mathbf{P}|\!|\!|_{\text{op}} \|\widehat{\Delta}\|_2^2 + 2|\langle \mathbf{U}^T \Delta,\, \underbrace{(\mathbf{U}^T \mathbf{P} \mathbf{U})}_{\widetilde{\mathbf{P}}}\, \mathbf{U}^T \theta^* \rangle|.$$

Noting that $\mathbf{U}^T \theta^* = e_1$ and hence $\widetilde{\mathbf{P}} \mathbf{U}^T \theta^* = \tilde{p}$, we have

$$|\langle \mathbf{U}^T \widehat{\Delta},\, \widetilde{\mathbf{P}}\, \mathbf{U}^T \theta^* \rangle| \leq \|\mathbf{U}^T \widehat{\Delta}\|_2 \, \|\tilde{p}\|_2 \;=\; \|\widehat{\Delta}\|_2 \, \|\tilde{p}\|_2.$$

Combining with the basic inequality (8.15), we have

$$\frac{\nu}{2}\|\widehat{\Delta}\|_2^2 \leq |||\mathbf{P}|||_{\text{op}}\, \|\widehat{\Delta}\|_2^2 + 2\|\widehat{\Delta}\|_2\, \|\tilde{p}\|_2.$$

Since $|||\mathbf{P}|||_{\text{op}} \leq \frac{\nu}{4}$ by assumption, we conclude that $\|\widehat{\Delta}\|_2 \leq \frac{8\|\tilde{p}\|_2}{\nu}$, as claimed.

We now turn to the remaining lemma.

Proof of Lemma 1: Since $\widehat{\theta}$ and θ^* are optimal and feasible respectively for the programs (8.14) with $\mathcal{C} = \mathbb{S}^{d-1}$, we are guaranteed that

$$\langle \theta^*,\, \widehat{\mathbf{\Sigma}}\, \theta^* \rangle \leq \langle \widehat{\theta},\, \widehat{\mathbf{\Sigma}} \widehat{\theta} \rangle. \tag{8.16}$$

Since both $\widehat{\theta}$ and $-\widehat{\theta}$ achieve this same maximum, we may assume without loss of generality that $\widehat{\theta}$ is chosen such that $\varrho := \langle \widehat{\theta},\, \theta^* \rangle \in [0, 1]$. Defining the matrix perturbation $\mathbf{P} = \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$, we have

$$\langle\langle \mathbf{\Sigma},\, \theta^* \otimes \theta^* - \widehat{\theta} \otimes \widehat{\theta} \rangle\rangle \leq \langle\langle \mathbf{P},\, \theta^* \otimes \theta^* - \widehat{\theta} \otimes \widehat{\theta} \rangle\rangle, \tag{8.17}$$

where $\langle\langle \mathbf{A},\, \mathbf{B} \rangle\rangle$ is the trace inner product, and $a \otimes a = aa^T$ is the rank one outer product. Following some simple algebra, the right-hand side is seen to be equal to $-\Psi(\widehat{\Delta}; \mathbf{P})$. The final step is to show that

$$\frac{\nu}{2}\|\widehat{\Delta}\|_2^2 \leq \langle\langle \mathbf{\Sigma},\, \theta^* \otimes \theta^* - \widehat{\theta} \otimes \widehat{\theta} \rangle\rangle. \tag{8.18}$$

Since $\|\widehat{\theta}\|_2 = \|\theta^*\|_2 = 1$, we have $\|\theta^* - \widehat{\theta}\|_2^2 = 2\big(1 - \langle \theta^*,\, \widehat{\theta} \rangle\big)$. Accordingly, we seek an expression in terms of $\varrho := \langle \theta^*,\, \widehat{\theta} \rangle$. Let us write $\widehat{\theta} = \varrho\theta^* + (\sqrt{1 - \varrho^2})\, z$, where the vector $z \in \mathbb{R}^d$ is orthogonal to θ^*. Define the matrix $\Gamma = \mathbf{\Sigma} - \gamma_1\big(\theta^* \otimes \theta^*\big)$, and note that $\mathbf{\Gamma}\theta^* = 0$ and $|||\mathbf{\Gamma}|||_{\text{op}} \leq \gamma_2$ by construction. Consequently, we can write

$$\begin{aligned}\langle\langle \mathbf{\Sigma},\, \theta^* \otimes \theta^* - \widehat{\theta} \otimes \widehat{\theta} \rangle\rangle &= \gamma_1 \langle\langle \theta^* \otimes \theta^*,\, \theta^* \otimes \theta^* - \widehat{\theta} \otimes \widehat{\theta} \rangle\rangle + \langle\langle \Gamma,\, \theta^* \otimes \theta^* - \widehat{\theta} \otimes \widehat{\theta} \rangle\rangle \\ &= (1 - \varrho^2)\big\{\gamma_1 - \langle\langle \Gamma,\, z \otimes z \rangle\rangle\big\}.\end{aligned}$$

Since $|||\mathbf{\Gamma}|||_{\text{op}} \leq \gamma_2$, we have $|\langle\langle \mathbf{\Gamma},\, z \otimes z \rangle\rangle| \leq \gamma_2$. Putting together the pieces, we have shown that

$$\langle\langle \mathbf{\Sigma},\, \theta^* \otimes \theta^* - \widehat{\theta} \otimes \widehat{\theta} \rangle\rangle \geq (1 - \varrho^2)\big\{\gamma_1 - \gamma_2\big\} \;=\; (1 - \varrho^2)\, \nu.$$

Since $\|\widehat{\Delta}\|_2^2 = 2(1 - \varrho) \leq 2(1 - \varrho^2)$, the claim (8.18) follows. □

■ 8.2.2 Consequences for principal components of a spiked ensemble

Theorem 8.1 applies to any form of matrix perturbation. In the context of principal component analysis, this perturbation takes a very specific form—namely, as the difference between the sample and population covariance matrices, More concretely, suppose that we drawn n i.i.d. samples $\{x_i\}_{i=1}^n$ from a zero-mean random vector with covariance $\mathbf{\Sigma}$. Principal component analysis is then based on the eigenstructure of the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$, and the goal is to draw conclusions about the eigenstructure of the population matrix.

In order to bring sharper focus to this issue, let us study how PCA behaves for a very simple class of covariance matrices, known as spiked covariance matrices. A sample $x_i \in \mathbb{R}^d$ from the *spiked covariance ensemble* takes the form

$$x_i \overset{\mathrm{d}}{=} \sqrt{\nu}\, \xi_i\, \theta^* + \sigma w_i \tag{8.19}$$

where $\xi_i \in \mathbb{R}$ is a zero-mean random variable with unit variance, and $w_i \in \mathbb{R}^d$ is a zero-mean random vector, independent of ξ_i, and with covariance matrix I. Thus, the random vector x_i is zero-mean with a covariance matrix of the form

$$\mathbf{\Sigma} := \nu\, \theta^*\, (\theta^*)^T + \sigma^2 \mathbf{I}_d. \tag{8.20}$$

By construction, for any $\nu > 0$, the vector θ^* is the unique maximal eigenvector of $\mathbf{\Sigma}$ with eigenvalue $\gamma_1(\mathbf{\Sigma}) = \nu + \sigma^2$. All other eigenvalues of $\mathbf{\Sigma}$ are located at σ^2, so that we have an eigengap $\gamma_1(\mathbf{\Sigma}) - \gamma_2(\mathbf{\Sigma}) = \nu$.

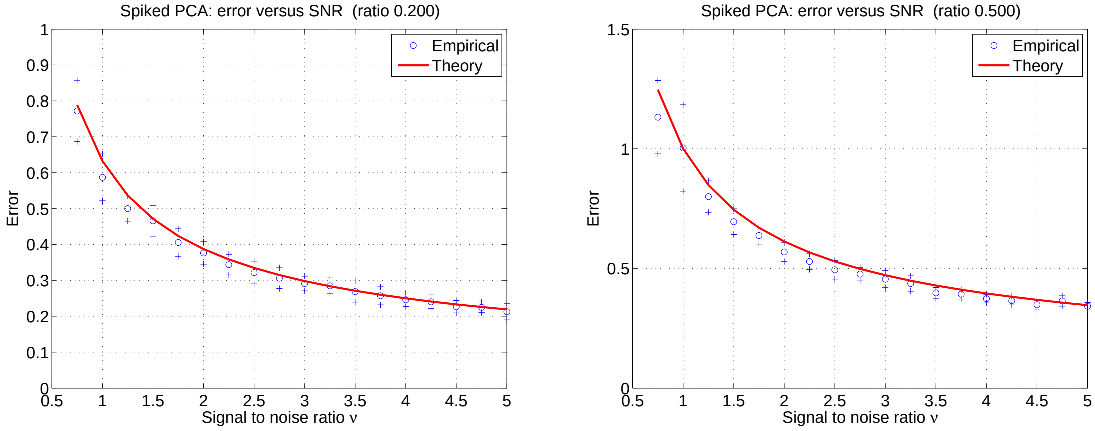
In the following result, we say that x_i has sub-Gaussian tails if both ξ_i and w_i are sub-Gaussian with parameter at most one.

Corollary 8.1. Given i.i.d. samples $\{x_i\}_{i=1}^n$ from the spiked covariance ensemble (8.19) with sub-Gaussian tails, and suppose that $n > d$ and $\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}} \leq \frac{1}{128}$. Then any maximal eigenvector $\widehat{\theta}$ of the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ satsifies the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq 32\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}} + \delta. \tag{8.21}$$

with probability at least $1 - c_1 e^{-c_2 n \min\{\sqrt{\nu}\delta,\, \nu\delta^2\}}$.

Figure 8-4 shows the results of simulations that confirm the qualitative scaling predicted by Corollary 8.1. In each case, we drew $n = 500$ samples from a spiked covariance matrix with the signal-to-noise parameter ν ranging over the interval $[0.75, 5]$. We then computed the ℓ_2-distance $\|\widehat{\theta} - \theta^*\|_2$ between the maximal eigenvectors of the sample and population covariances respectively, performing $T = 20$ trials for each



(a) (b)

Figure 8-4. Plots of the error $\|\widehat{\theta} - \theta^*\|_2$ versus the signal-to-noise ratio, as measured by the eigengap ν. Both plots are based on a sample size $n = 500$. Blue dots show the average of 100 trials, along with the standard errors (blue crosses). The red curve shows the theoretical bound $\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}}$. (a) Dimension $d = 100$. (b) Dimension $d = 250$.

setting of ν. The blue curves in Figure 8-4 show the empirical results, whereas the red curve corresponds to the theoretical prediction $\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}}$. Note that Corollary 8.1 predicts this scaling, but with a much poorer leading constant. As shown by Figure 8-4, Corollary 8.1 accurately captures the scaling behavior of the error as a function of the SNR.

Proof. Let $\mathbf{P} = \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}$ be the difference between the sample and population covariance matrices. In order to apply Theorem 8.1, we need to upper bound the quantities $|\!|\!|\mathbf{P}|\!|\!|_{\text{op}}$ and $\|\tilde{p}\|_2$. Defining the random vector $\bar{w} := \frac{1}{n}\sum_{i=1}^n \xi_i w_i$, the perturbation matrix $\mathbf{P}$ can be decomposed as

$$\mathbf{P} = \underbrace{\nu\big(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - 1\big)\theta^*(\theta^*)^T}_{\mathbf{P}_1} + \underbrace{\sqrt{\nu}\big(\bar{w}(\theta^*)^T + \theta^*\bar{w}^T\big)}_{\mathbf{P}_2} + \underbrace{\big(\frac{1}{n}\sum_{i=1}^n w_i w_i^T - I\big)}_{\mathbf{P}_3} \tag{8.22}$$

Since $\|\theta^*\|_2 = 1$, we thus have the upper bound

$$|\!|\!|\mathbf{P}|\!|\!|_{\text{op}} \leq \nu\Big|\frac{1}{n}\sum_{i=1}^n \xi_i^2 - 1\Big| + 2\sqrt{\nu}\|\bar{w}\|_2 + |\!|\!|\frac{1}{n}\sum_{i=1}^n w_i w_i^T - I|\!|\!|_{\text{op}}. \tag{8.23}$$

Let us derive a similar upper bound on $\|\tilde{p}\|_2$ using the decomposition (8.11). Since θ^*

is the unique maximal eigenvector of $\mathbf{\Sigma}$, it forms the first column of the matrix $\mathbf{U}$. Let $\mathbf{U}_2 \in \mathbb{R}^{d \times (d-1)}$ denote the matrix formed of the remaining $(d-1)$ columns. With this notation, we have $\tilde{p} = \mathbf{U}_2^T \mathbf{P} \theta^*$. Using the decomposition (8.22) of the perturbation matrix and the fact that $\mathbf{U}_2^T \theta^* = 0$, we find that $\tilde{p} = \sqrt{\nu}\, \mathbf{U}_2^T \bar{w} + \frac{1}{n} \sum_{i=1}^n \mathbf{U}_2^T\, w_i \, \langle w_i,\, \theta^* \rangle$. Since $\mathbf{U}_2$ has orthonormal columns, we have $\|\mathbf{U}_2^T \bar{w}\|_2 \leq \|\bar{w}\|_2$ and also

$$\|\sum_{i=1}^n \mathbf{U}_2^T \, w_i \, \langle w_i,\, \theta^* \rangle\|_2 = \sup_{\|v\|_2 = 1} \big|(\mathbf{U}_2 v)^T \big(\sum_{i=1}^n w_i w_i^T - \mathbf{I}_d\big) \theta^* \big| \;\leq\; |\!|\!| \frac{1}{n} \sum_{i=1}^n w_i w_i^T - \mathbf{I}_d |\!|\!|_{\mathrm{op}}.$$

Putting together the pieces, we have shown that

$$\|\tilde{p}\|_2 \leq \sqrt{\nu}\, \|\bar{w}\|_2 + |\!|\!| \frac{1}{n} \sum_{i=1}^n w_i w_i^T - \mathbf{I}_d |\!|\!|_{\mathrm{op}}. \tag{8.24}$$

The following lemma allows us to control the quantities appearing the bounds (8.23) and (8.24):

Lemma 2. *Under the conditions of Corollary 8.1, we have*

$$\mathbb{P}\big[\big|\frac{1}{n} \sum_{i=1}^n \xi_i^2 - 1\big| \geq \delta_1\big] \leq 2 e^{-c_2 n \min\{\delta_1,\, \delta_1^2\}}, \tag{8.25a}$$

$$\mathbb{P}\big[\|\bar{w}\|_2 \geq 2\sqrt{\frac{d}{n}} + \delta_2\big] \leq 2 e^{-c_2 n \min\{\delta_2,\, \delta_2^2\}}, \quad \textit{and} \tag{8.25b}$$

$$\mathbb{P}\Big[|\!|\!| \frac{1}{n} \sum_{i=1}^n w_i w_i^T - \mathbf{I}_d |\!|\!|_{op} \geq 2\sqrt{\frac{d}{n}} + 2\delta_3 + \big(\sqrt{\frac{d}{n}} + \delta_3\big)^2\Big] \leq 2 e^{-n\delta_3^2/2}. \tag{8.25c}$$

We leave this proof as an exercise, since it is straightforward application of results and techniques from previous chapters. For future reference, we define

$$\phi(\delta_1, \delta_2, \delta_3) := 2 e^{-c_2 n \min\{\delta_1,\, \delta_1^2\}} + 2 e^{-c_2 n \min\{\delta_2,\, \delta_2^2\}} + 2 e^{-n \delta_3^2/2}, \tag{8.26}$$

corresponding to the probability with which one of the bounds in Lemma 2 is violated.

In order to apply Theorem 8.1, we need to first show that $|\!|\!|\mathbf{P}|\!|\!|_{\mathrm{op}} \leq \frac{\nu}{4}$ with high probability. Beginning with the inequality (8.23) and applying Lemma 2 with $\delta_1 = \frac{1}{16}$, $\delta_2 = \frac{\delta}{4\sqrt{\nu}}$ and $\delta_3 = \delta/16 \in (0, 1)$, we have

$$|\!|\!|\mathbf{P}|\!|\!|_{\mathrm{op}} \leq \frac{\nu}{16} + 8(\sqrt{\nu} + 1) \sqrt{\frac{d}{n}} + \delta \;\leq\; \frac{\nu}{16} + 16(\sqrt{\nu + 1}) \sqrt{\frac{d}{n}} + \delta$$

with probability at least $1 - \phi\big(\frac{1}{4}, \frac{\delta}{3\sqrt{\nu}}, \frac{\delta}{16}\big)$. Consequently, as long $\sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{d}{n}} \leq \frac{1}{128}$, we have

$$|\!|\!|\mathbf{P}|\!|\!|_{\text{op}} \le \frac{3}{16}\nu + \delta < \frac{\nu}{4}, \quad \text{for all } \delta \in \tfrac{\nu}{16}.$$

It remains to bound $\|\tilde{p}\|_2$. Applying Lemma 2 to the inequality (8.24) with the previously specified choices of $(\delta_1, \delta_2, \delta_3)$, we have

$$\|\tilde{p}\|_2 \le 2\big(\sqrt{\nu}+1\big)\sqrt{\frac{d}{n}} + \delta \;\le\; 4\sqrt{\nu+1}\sqrt{\frac{d}{n}} + \delta$$

with probability at least $1 - \phi\big(\frac{1}{4}, \frac{\delta}{3\sqrt{\nu}}, \frac{\delta}{16}\big)$. We have shown that conditions of Theorem 8.1 are satisfied, so that the claim (8.21) follows as a consequence of the bound (8.12).

□

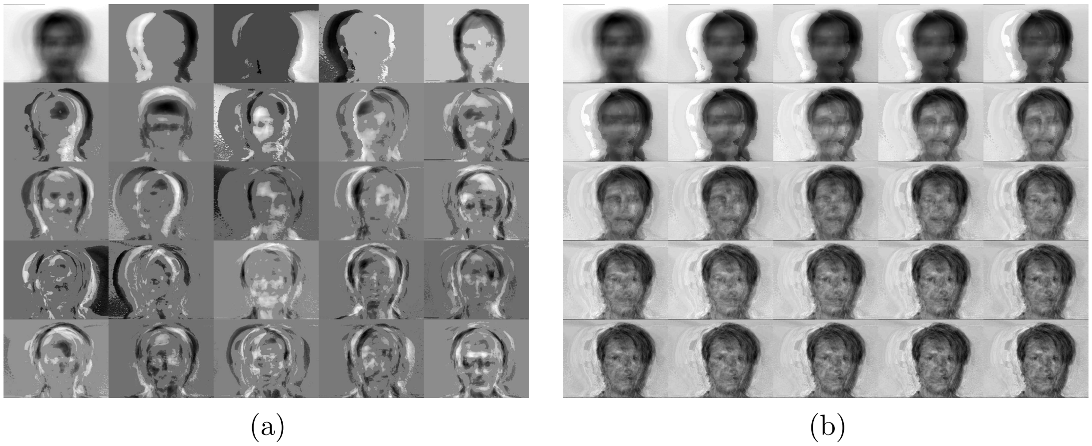
■ 8.3 Sparse PCA

Corollary 8.1 hints that ordinary PCA does not behave well when the ratio d/n is large relative to the eigengap. In practice, however, it is often reasonable to impose structure on eigenvectors, and this structure can be exploited. Perhaps the simplest such structure is that of sparsity. Accordingly, this section is devoted to the sparse version of principal component analysis.

What are some reasons for studying the problem of sparse PCA? As mentioned above, one reason is that classical PCA behaves poorly when the ratio d/n does not converge to zero. Indeed, as discussed at more length in the bibliographic section, if the ratio d/n stays suitably bounded away from zero, as a function of the signal-to-noise ratio, then the eigenvectors of the sample covariance in a spiked covariance model become asymptotically orthogonal to their population analogues. One might ask whether the population eigenvectors might be estimated consistently using a method more sophisticated than PCA. This question has a negative answer: as we discuss in Chapter 15, for the standard spiked model (8.19), no method can produce consistent estimators of the population eigenvectors when d/n stays bounded away from zero.

A second reason—valid even when there is no reason to believe that the true eigenvectors are sparse—is that sparse eigenvalues may be more interpretable than their dense analogues. Let us illustrate by revisiting the eigenfaces from Example 8.2.

Example 8.5 (Sparse eigenfaces)**.** Using the same images from the Yale Face Database, we approximated the sparse eigenvectors with sparsity $s = \lfloor 0.25d \rfloor = 19440$ dimensions. In order to do so, we applied a thresholded version of the matrix power method for computing eigenvalues and eigenvectors. (See Exercise 8.5 for exploration of the standard matrix power method.) Panel (a) of Figure 8-5 shows the average face (top left image), along with approximations to the first 24 sparse eigenfaces. Each sparse eigenface was


(a)

(b)

Figure 8-5. Illustration of sparse eigenanalysis for the Yale Face Database. (a) Average face (top left image), and approximations to the first 24 sparse eigenfaces, obtained by a greedy iterative thresholding procedure applied to the eigenvalue power method. Eigenfaces were restricted to have at most 25% of their pixels non-zero, corresponding to a 1/4 reduction in storage. (b) Reconstruction based on sparse eigenfaces.

restricted to have at most 25% of its pixels non-zero, corresponding to a savings of a factor of 4 in storage. Note that the sparse eigenfaces are more localized than their PCA analogues from Figure 8-2. Panel (b) shows reconstruction using the average face in conjunction with the first 100 sparse eigenfaces, which require equivalent storage (in terms of pixel values) to the first 25 regular eigenfaces. ♣

■ 8.3.1 A general deterministic result

We now turn to the question of how to estimate a maximal eigenvector that is known *a priori* to be sparse. A natural approach is to augment the quadratic objective function which underlies classical PCA with an additional sparsity constraint or penalty. More concretely, we analyze both the the constrained problem

$$\widehat{\theta} \in \arg\max_{\|\theta\|_2=1} \big\{\langle \theta, \, \widehat{\mathbf{\Sigma}}\,\theta\rangle\big\} \qquad \text{such that } \|\theta\|_1 \le R, \tag{8.27}$$

as well as the penalized variant

$$\widehat{\theta} \in \arg\max_{\|\theta\|_2=1} \big\{\langle \theta, \, \widehat{\mathbf{\Sigma}}\,\theta\rangle - \Phi_{\lambda_n}(\theta)\big\}, \quad \text{where} \quad \Phi_{\lambda_n}(\theta) := \lambda_n \|\theta\|_1 + \lambda_n^2 \|\theta\|_1^2. \tag{8.28}$$

Here the matrix $\widehat{\mathbf{\Sigma}}$ represents some type of approximation to the population covariance matrix $\mathbf{\Sigma}$, with the sample covariance being a canonical example. Note that neither estimator is convex, since they involve maximization of a positive semidefinite quadratic form over a non-convex constraint set. Nonetheless, it is instructive to analyze them in order to understand the statistical behavior of sparse PCA, and in the exercises, we describe some relaxations of these nonconvex programs.

Naturally, the proximity of $\widehat{\theta}$ to the maximum eigenvector θ^* of $\mathbf{\Sigma}$ depends on the perturbation matrix $\mathbf{P} := \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$. How to measure the effect of the perturbation? As will become clear, much of our analysis of ordinary PCA can be modified in a relatively straightforward way so as to obtain results for the sparse version. In particular, a central object in our analysis of ordinary PCA was the basic inequality stated in Lemma 1: it shows that perturbation matrix enters via the function

$$\Psi(\Delta; \mathbf{P}) := \langle \Delta, \, \mathbf{P}\Delta \rangle + 2\langle \Delta, \, \mathbf{P}\theta^* \rangle.$$

As with our analysis of PCA, our general deterministic theorem for sparse PCA involves imposing a form of uniform control on $\Psi(\Delta; \mathbf{P})$ as Δ ranges over all vectors of the form $\theta - \theta^*$ with $\theta \in \mathbb{S}^{d-1}$. The sparsity constraint enters in the form of this uniform bound that we assume. More precisely, letting $\varphi(n,d)$ be a non-negative function of the sample size and dimension, we assume that there exist universal constants c_0 and c_1 such that

$$\sup_{\substack{\Delta = \theta - \theta^* \\ \|\theta\|_2 = 1}} \big|\Psi(\Delta; \mathbf{P})\big| \leq c_0 \, \nu \, \|\Delta\|_2^2 + c_1 \Big\{ \big(\sqrt{(\nu+1)}\, \varphi(n,d) \|\Delta\|_1 + \varphi^2(n,d) \|\Delta\|_1^2 \Big\}, \qquad (8.29)$$

As a concrete example, for a sparse version of the spiked PCA ensemble (8.19) with sub-Gaussian tails, this condition is satisfied with high probability with $\varphi(n,d) = \sqrt{\frac{\log d}{n}}$. This fact will be established in the proof of Corollary 8.2 to follow.

Theorem 8.2. Let $\widehat{\mathbf{\Sigma}}$ be any symmetric matrix satisfying condition (8.29) with constant $c_0 < \frac{1}{4}$, and suppose that $9 c_1 s \, \varphi^2(n,d) \leq \frac{1}{8}$.

(a) For any optimal solution $\widehat{\theta}$ to the constrained program (8.27) with $R = \|\theta^*\|_1$:

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{3c_1}{1/4 - c_0} \sqrt{\frac{\nu+1}{\nu^2}} \, \sqrt{s} \, \varphi(n,d) \qquad (8.30)$$

(b) Given a regularization parameter $\lambda_n \geq \big(1 + \frac{3c_1}{4}\big)\sqrt{\nu+1}\, \varphi(n,d)$, suppose that the sample size is sufficiently large so as to ensure that $3\lambda_n^2 s < \frac{1}{8}$. Then any optimal solution $\widehat{\theta}$ to the regularized program (8.28) satisfies the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{1}{1/4 - c_0} \left\{ \frac{\lambda_n}{\nu} + 3c_1 \sqrt{\frac{\nu+1}{\nu^2}} \right\} \sqrt{s} \, \varphi(n,d) \qquad (8.31)$$

Proof. We begin by analyzing the constrained estimator, and then describe the modifications necessary for the regularized version.

Argument for constrained estimator: Note that $\|\widehat{\theta}\|_1 \leq R = \|\theta^*\|_1$ by construction of the estimator, and moreover $\theta^*_{S^c} = 0$ by assumption. By splitting the ℓ_1-norm into two components, indexed by S and S^c respectively, it can be shown[2] that the error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ satisfies the inequality $\|\widehat{\Delta}_{S^c}\|_1 \leq \|\widehat{\Delta}_S\|_1$. So as to simplify our treatment of the regularized estimator, let us proceed assuming only the (weaker) inequality $\|\widehat{\Delta}_{S^c}\|_1 \leq 2\|\widehat{\Delta}_S\|_1$, which implies that $\|\widehat{\Delta}\|_1 \leq 3\sqrt{s}\|\widehat{\Delta}\|_2$. Combining this inequality with the condition (8.29), we find that

$$\big|\Psi(\widehat{\Delta}; \mathbf{P})\big| \leq c_0\, \nu\, \|\widehat{\Delta}\|_2^2 + c_1 \Big\{ 3\sqrt{\nu+1}\, \sqrt{s}\varphi(n,d) \|\widehat{\Delta}\|_2 + 9s\, \varphi^2(n,d) \|\widehat{\Delta}\|_2^2 \Big\} \tag{8.32}$$

Substituting back into the basic inequality (8.15) and performing some algebra yields

$$\underbrace{\Big\{\frac{1}{2} - c_0 - 9c_1 s\, \varphi^2(n,d)\Big\}}_{\kappa} \nu \|\widehat{\Delta}\|_2^2 \leq 3c_1 \sqrt{\frac{\nu+1}{\nu^2}}\, \sqrt{\frac{s\log d}{n}} \nu \|\widehat{\Delta}\|_2.$$

Note that our assumptions imply that $\kappa > \frac{1}{4} - c_0$, so that the bound (8.30) follows.

Argument for regularized estimator: We now turn to the regularized estimator (8.28). With the addition of the regularizer, the basic inequality (8.15) now takes the slightly modified form

$$\frac{\nu}{2}\|\widehat{\Delta}\|_2^2 \leq |\Psi(\widehat{\Delta}; \mathbf{P})| + \lambda_n \big\{\Phi_{\lambda_n}(\theta^*) - \Phi_{\lambda_n}(\widehat{\theta})\big\}. \tag{8.33}$$

We need to use the term involving Φ_{λ_n} to prove that the error vector $\widehat{\Delta}$ still satisfies the cone inequality $\|\widehat{\Delta}\|_1 \leq 3\sqrt{s}\|\widehat{\Delta}\|_2$. We state this claim as a separate lemma.

Lemma 3. *Under the conditions of Theorem 8.2, the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ satisfies the inequality*

$$\|\widehat{\Delta}_{S^c}\|_1 \leq 2\|\widehat{\Delta}_S\|_1 \quad \textit{and hence,} \quad \|\widehat{\Delta}\|_1 \leq 3\sqrt{s}\|\widehat{\Delta}\|_2. \tag{8.34}$$

Taking this lemma as given, let us complete the proof of the theorem. Given Lemma 3, the previously derived upper bound (8.32) on $|\Psi(\widehat{\Delta}; \mathbf{P})|$ is also applicable to the regularized estimator. The only remaining detail is to upper bound the additional terms in our basic inequality (8.35). In particular, we have

$$\lambda_n \big\{\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\big\}\ \big\{1 + \lambda_n\|\widehat{\Delta}\|_1\big\} \leq \lambda_n\sqrt{s}\|\widehat{\Delta}\|_2 + 3\lambda_n^2 s\|\widehat{\Delta}\|_2^2.$$

[2] We leave this calculation as an exercise for the reader; otherwise, helpful details can be found in Chapter 7.

Together with our previous bound (8.32), we obtain

$$\underbrace{\Big\{\frac{1}{2} - c_0 - 9c_1 s\,\varphi^2(n,d) - 3\lambda_n^2 s\Big\}}_{\kappa'}\,\nu\|\widehat{\Delta}\|_2^2 \le \Big\{\lambda_n + 3c_1\sqrt{\nu+1}\,\Big\}\,\sqrt{s}\varphi(n,d)$$

Our choice of λ_n and assumption on n ensure that $\kappa' \geq \frac{1}{4} - c_0$, from which the claim (8.31) follows.

It remains to prove Lemma 3. By separating the ℓ_1-norm into components indexed by S and S^c respectively, it can be verified that

$$\|\theta^*\|_1 - \|\widehat{\theta}\|_1 \le \|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1, \quad \text{and} \quad \|\theta^*\|_1^2 - \|\widehat{\theta}\|_1^2 \le \big\{\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\big\}\,\|\widehat{\Delta}\|_1.$$

Substituting into the basic inequality (8.33), we find that

$$\frac{\nu}{2}\|\widehat{\Delta}\|_2^2 \le |\Psi(\widehat{\Delta};\mathbf{P})| + \lambda_n\big\{\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\big\}\,\big\{1 + \lambda_n\|\widehat{\Delta}\|_1\big\}. \tag{8.35}$$

Proceeding via proof by contradiction, suppose that $\|\widehat{\Delta}_S\|_1 < \frac{1}{2}\|\widehat{\Delta}_{S^c}\|_1$. We then have $\|\widehat{\Delta}_{S^c}\|_1 \geq \frac{2}{3}\|\widehat{\Delta}\|_1$, and hence that

$$\lambda_n\big\{\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\big\}\,\big\{1 + \lambda_n\|\widehat{\Delta}\|_1\big\} \le -\frac{4\lambda_n}{3}\|\widehat{\Delta}\|_1 - \frac{4{\lambda_n}^2}{3}\|\widehat{\Delta}\|_1^2.$$

We substitute this inequality together with the assumed bound (8.29) on Ψ into the basic inequality (8.15), thereby obtaining

$$\frac{\nu}{2}\|\widehat{\Delta}\|_2^2 \le c_0\nu\|\widehat{\Delta}\|_2^2 + \Big\{c_1\sqrt{\nu+1}\varphi(n,d) - \frac{4\lambda_n}{3}\Big\}\|\widehat{\Delta}\|_1 + \Big\{c_1\varphi^2(n,d) - \frac{4\lambda_n^2}{3}\Big\}\|\widehat{\Delta}\|_1^2.$$

Our assumption on λ_n ensures that both of the terms involving the ℓ_1-norm are non-positive, and hence we obtain the inequality $(\frac{1}{2} - c_0)\nu\|\widehat{\Delta}\|_2^2 \le 0$. Since $c_0 < 1/2$, this inequality implies that $\widehat{\Delta} = 0$, which contradicts the assumption that $\|\widehat{\Delta}_S\|_1 < \frac{1}{2}\|\widehat{\Delta}_{S^c}\|_1$.

□

■ 8.3.2 Consequences for the spiked model with sparsity

Theorem 8.2 is a general deterministic guarantee that applies to any matrix with a sparse maximal eigenvector. In order to obtain more concrete results in a particular case, let us return to the spiked covariance model previously introduced in equation (8.19), and analyze a sparse variant of it. More precisely, consider a random vector

$x_i \in \mathbb{R}^d$ generated as

$$x_i \overset{\mathrm{d}}{=} \sqrt{\nu}\,\xi_i \theta^* + w_i, \tag{8.36}$$

where $\theta^* \in \mathbb{S}^{d-1}$ is a s-sparse vector, corresponding to the maximal eigenvector of $\mathbf{\Sigma} = \operatorname{cov}(x_i)$. As before, we assume that both the random variable ξ_i and the random vector $w_i \in \mathbb{R}^d$ are sub-Gaussian with parameter 1, in which case we say that the random vector $x_i \in \mathbb{R}^d$ has sub-Gaussian tails.

Corollary 8.2. Consider n i.i.d. samples $\{x_i\}_{i=1}^n$ from a s-sparse spiked covariance matrix with eigengap $\nu > 0$ and suppose that $n > \frac{c}{1+\nu^2}\, s \log d$ for a sufficiently large constant c. Then any optimal solution $\widehat{\theta}$ to the program (8.27) with $R = \|\theta^*\|_1$, or to the program (8.28) with $\lambda_n^2 = 8(\nu+1)\frac{\log d}{n}$ satisfies the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq c_0 \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{s \log d}{n}} + \delta \qquad \text{for all } \delta \in (0,1) \tag{8.37}$$

with probability at least $1 - c_1 e^{-c_2 n \min\{\delta^2, \nu^2\}}$.

Proof. Letting $\mathbf{P} = \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$ be the deviation between the sample and population covariance matrices, our goal is to show that $\mathbf{P}$ satisfies the condition (8.29). Recall from equation (8.22) the decomposition $\mathbf{P} = \sum_{j=1}^3 \mathbf{P}_j$: by linearity of the function Ψ, this decomposition implies that $\Psi(\Delta; \mathbf{P}) = \sum_{j=1}^3 \Psi(\Delta; \mathbf{P}_j)$. We control each of these terms in turn.

Beginning with $\Psi(\Delta; \mathbf{P}_1)$, Lemma 2 guarantees that $\left|\frac{1}{n}\sum_{i=1}^n \xi_i^2 - 1\right| \leq \frac{1}{16}$ with probability at least $1 - 2e^{-cn}$. Conditioned on this bound, for any vector of the form $\Delta = \theta - \theta^*$ with $\theta \in \mathbb{S}^{d-1}$, we have

$$|\Psi(\Delta; \mathbf{P}_1)| \leq \frac{\nu}{16} \langle \Delta,\, \theta^* \rangle^2 \;=\; \frac{\nu}{16}\big(1 - \langle \theta^*,\, \theta\rangle\big)^2 \;\leq\; \frac{\nu}{32}\|\Delta\|_2^2, \tag{8.38}$$

where we have used the fact that

$$2\big(1 - \langle \theta^*,\, \theta\rangle\big)^2 \leq 2\big(1 - \langle \theta^*,\, \theta\rangle\big) \;=\; \|\Delta\|_2^2.$$

Turning to the second term, we have

$$\begin{aligned} |\Psi(\Delta; \mathbf{P}_2)| &\leq 2\sqrt{\nu}\Big\{\langle \Delta,\, \bar{w}\rangle \langle \Delta,\, \theta^*\rangle + \langle \bar{w},\, \Delta\rangle + \langle \theta^*,\, \bar{w}\rangle\langle \Delta,\, \theta^*\rangle\Big\} \\ &\leq 4\sqrt{\nu}\|\Delta\|_1 \|\bar{w}\|_\infty + 2\sqrt{\nu}|\langle \theta^*,\, \bar{w}\rangle| \frac{\|\Delta\|_2^2}{2} \end{aligned} \tag{8.39}$$

The following lemma provides control on the two terms in this upper bound:

Lemma 4. *Under the conditions of Corollary 8.2, we have*

$$\mathbb{P}\big[\|\bar{w}\|_\infty \geq 4\sqrt{\frac{\log d}{n}} + \delta\big] \leq c_1 e^{-c_2 n \delta^2} \quad \textit{for all } \delta \in (0,1), \textit{ and} \tag{8.40a}$$

$$\mathbb{P}\big[|\langle \theta^*, \, \bar{w}\rangle| \geq \frac{\sqrt{\nu}}{16}\big] \leq c_1 e^{-c_2 n \, \nu}. \tag{8.40b}$$

We leave the proof of these bounds as an exercise for the reader, since they follow from standard results in Chapter 2. Combining Lemma 4 with the bound (8.39) yields

$$|\Psi(\Delta; \mathbf{P}_2)| \leq 16\Big\{\sqrt{\frac{(\nu+1)\, \log d}{n}} + \delta\Big\}\|\Delta\|_1 + \frac{\nu}{16}\|\Delta\|_2^2 \tag{8.41}$$

Turning to the final term involving $\mathbf{P}_3 = \frac{1}{n}\mathbf{W}^T\mathbf{W} - \mathbf{I}_d$, we have

$$|\Psi(\Delta; \mathbf{P}_3)| \leq \big|\langle \Delta, \, \mathbf{P}_3 \Delta\rangle\big| + 2\|\mathbf{P}_3\theta^*\|_\infty \|\Delta\|_1. \tag{8.42}$$

Our final lemma controls the two terms in this bound:

Lemma 5. *Under the conditions of Corollary 8.2, for all $\delta \in (0,1)$, we have*

$$\|\mathbf{P}_3 \theta^*\|_\infty \leq 4\sqrt{\frac{\log d}{n}} + \delta/2, \quad \textit{and} \tag{8.43a}$$

$$\sup_{\Delta \in \mathbb{R}^d} \big|\langle \Delta, \, \mathbf{P}_3 \Delta\rangle\big| \leq \frac{\nu}{16}\|\Delta\|_2^2 + \frac{1}{16}\frac{\log d}{n}\|\Delta\|_1^2 \tag{8.43b}$$

where both inequalities hold with probability greater than $1 - c_1 e^{-c_2 n \min\{\nu^2, \delta^2\}}$.

Combining this lemma with our earlier inequality (8.42) yields the bound

$$|\Psi(\Delta; \mathbf{P}_3)| \leq \frac{\nu}{16}\|\Delta\|_2^2 + \frac{1}{16}\frac{\log d}{n}\|\Delta\|_1^2 + 8\big(\sqrt{\frac{\log d}{n}} + \delta\big)\|\Delta\|_1 + \delta. \tag{8.44}$$

Finally, combining the bounds (8.38), (8.41) and (8.44), we find that, for all $\Delta \in \mathbb{R}^d$,

$$\big|\Psi(\Delta; \mathbf{P})\big| \leq \frac{\nu}{8}\|\Delta\|_2^2 + 24\Big\{\sqrt{\frac{(\nu+1)\log d}{n}} + \delta\Big\}\|\Delta\|_1 + \frac{1}{16}\frac{\log d}{n}\|\Delta\|_1^2 + \delta$$

with probability at least $1 - c_1 e^{-c_2 n \min\{\delta^2, \, \nu^2\}}$.

The only remaining detail is the proof of Lemma 5. The proof of the bound (8.43a) is a simple exercise, using the sub-exponential tail bounds from Chapter 2. The proof of the bound (8.43b) requires more involved argument, one that makes use of both Exercise 7.10 and our previous results on estimation of sample covariances from Chapter 6.

Recall that $\mathbf{P}_3 = \frac{1}{n}\sum_{i=1}^n w_i w_i^T - \mathbf{I}_d$. Introducing the shorthand $k = \max\{1, \lceil \frac{n}{\log(d)} \rceil\}$, consider the sub-matrices $\{(\mathbf{P}_3)_{SS}, |S| = k\}$. Given a parameter $\alpha \in (0,1)$ to be chosen, a combination of the union bound and Theorem 6.2 imply that there are universal constants c_1 and c_2 such that

$$\mathbb{P}\Big[\max_{|S|=k} |\!|\!|(\mathbf{P}_3)_{SS}|\!|\!|_{\mathrm{op}} \geq c_1\sqrt{\frac{k}{n}} + \alpha\min\{1,\nu\}\Big] \leq 2e^{-c_2\alpha^2 n \min\{1,\nu^2\} + \log\binom{d}{k}}$$

Since $\log\binom{d}{k} \leq k\log(ed)$, our choice of k and the assumption that $n > \frac{c}{1+\nu^2}\log d$ for a suitably large constant c implies that

$$\mathbb{P}\Big[\max_{|S|=k} |\!|\!|(\mathbf{P}_3)_{SS}|\!|\!|_{\mathrm{op}} \geq c_1'\alpha\min\{1,\nu\}\Big] \leq 2e^{-c_2'\alpha^2 n \min\{1,\nu^2\}}.$$

The result of Exercise 7.10 then implies that

$$\big|\langle \Delta,\, \mathbf{P}_3\Delta\rangle\big| \leq 27c_1'\alpha\min\{1,\nu\}\Big\{\|\Delta\|_2^2 + \frac{\log d}{n}\|\Delta\|_1^2\Big\} \qquad \text{for all } \Delta \in \mathbb{R}^d,$$

with the previously stated probability. Setting $\alpha = \frac{1}{(16\times 27)c_1'}$ yields the claim (8.43b).

■ 8.4 Bibliographic details and background

Further details on PCA and its applications can be found in books by Jolliffe [Jol04] and Anderson [And84] (cf. Chapter 11). The two-volume set by Horn and Johnson [HJ85, HJ91] contains a wealth of background on matrix analysis; see also the book by Bhatia [Bha97] for a general operator-theoretic viewpoint. The book by Stewart and Sun [SS80] is more specifically focused on matrix perturbation theory, whereas Stewart [Ste71] provides perturbation theory for closed linear operators.

Johnstone [Joh01] introduced the spiked covariance model (8.19), and investigated the high-dimensional asymptotics of its eigenstructure. Johnstone and Lu [JL09] introduced the sparse variant of the spiked ensemble, and proved consistency results for a simple estimator based on thresholding the diagonal entries of the sample covariance matrix. Amini and Wainwright [AW09] provided a more refined analysis of this same estimator, as well as of a semidefinite programming (SDP) relaxation proposed by d'Asprémont et al. [dEJL07]. See Exercise 8.9 for the derivation of this latter SDP relaxation. The nonconvex estimator (8.27) was first proposed by Joliffe et al. [JTU03], and called the `SCOTLASS` criterion; Witten et al. [WTH09] derive an alternating algorithm for finding a local optimum of this criterion. Other authors, including Ma [Ma10, Ma13] and Yuan and Zhang [YZ13], have studied iterative algorithms for sparse PCA based on combining the power method with soft thresholding.

Minimax lower bounds on variable selection for sparse PCA were derived by Amini and Wainwright [AW09], whereas lower bounds for estimation in ℓ_2 and related norms were derived by Birnbaum et al. [BJNP12], and Vu and Lei [VL12]. These types of lower bounds are based on the minimax theory covered in Chapter 15. Berthet and Rigollet [BR13] derived certain hardness results for the problem of sparse PCA detection, based on relating it to the (conjectured) average-case hardness of the planted k-clique problem in Erdös-Renyi random graphs. See also Ma and Wu [MW13] for a slightly different reduction to the k-clique problem, one which applies to a detection problem over a family of sparse-plus-low-rank matrices. Since estimation is in general harder than detection, these results also imply computational lower bounds on estimation in sparse PCA, again conditional on the average-case hardness of the k-clique problem.

■ 8.5 Exercises

Exercise 8.1 (Courant-Fischer variational representation)**.** For a given $j \in \{2, \ldots, d\}$, let $\mathcal{V}_{j-1}$ denote the collection of all subspaces of dimension $j-1$. For any symmetric matrix $\mathbf{Q}$, show that the j^{th} largest eigenvalue is given by

$$\gamma_j(\mathbf{Q}) = \min_{\mathbb{V} \in \mathcal{V}_{j-1}} \max_{u \in \mathbb{V}^\perp \cap \mathbb{S}^{d-1}} \langle u, \, \mathbf{Q} u \rangle, \tag{8.45}$$

where $\mathbb{V}^\perp$ denotes the orthogonal subspace to $\mathbb{V}$.

Exercise 8.2. For positive integers $d_1 \leq d_2$, a matrix norm on $\mathbb{R}^{d_1 \times d_2}$ is *unitarily invariant* if $|||\mathbf{M}||| = |||\mathbf{V}\mathbf{M}\mathbf{U}|||$ for all orthonormal matrices $\mathbf{V} \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{U} \in \mathbb{R}^{d_2 \times d_2}$.

(a) Prove that the Frobenius norm $|||\mathbf{M}|||_{\mathrm{F}} = \sqrt{\sum_{j=1}^{d_1} \sum_{k=1}^{d_2} M_{jk}^2}$ is unitarily invariant.

(b) Let ρ be a norm on $\mathbb{R}^{d_1}$ that is invariant to permutations and sign changes—that is

$$\rho(x_1, \ldots, x_{d_1}) = \rho\big(z_1 x_{\pi(1)}, \ldots, z_{d_1} x_{\pi(d_1)}\big)$$

for all binary strings $z \in \{-1, 1\}^{d_1}$ and permutations π on $\{1, \ldots, d_1\}$. Such a function is known as a *symmetric gauge function.* Letting $\{\gamma_j(\mathbf{M})\}_{j=1}^{d_1}$ denote the singular values of $\mathbf{M}$, show that

$$|||\mathbf{M}|||_\star := \rho\big(\gamma_1(\mathbf{M}), \ldots, \gamma_{d_1}(\mathbf{M})\big).$$

defines a matrix norm.

(c) Show that all matrix norms in the family from part (b) are unitarily invariant.

(d) Prove that the following norms are unitarially invariant:

(i) the Frobenium norm $|||\mathbf{M}|||_{\mathrm{F}}$

(ii) the nuclear norm $|||\mathbf{M}|||_{\mathrm{nuc}}$

(iii) the ℓ_2-operator norm $|||\mathbf{M}|||_{\mathrm{op}}$.

Exercise 8.3. Prove Weyl's inequality (8.9). (*Hint:* Exercise 8.1 may be useful.)

Exercise 8.4. Show that the orthogonal matrix $\mathbf{V} \in \mathbb{R}^{d \times r}$ maximizing the criterion (8.2) has the top r eigenvectors of the covariance matrix $\mathbf{\Sigma}$ as its columns.

Exercise 8.5 (Matrix power method)**.** Let $\mathbf{Q} \in \mathcal{S}^{d \times d}$ be a strictly positive definite symmetric matrix with a unique maximal eigenvector θ^*. Given some non-zero initial vector $\theta^0 \in \mathbb{R}^d$, consider the sequence $\{\theta^t\}_{t=0}^{\infty}$

$$\theta^{t+1} = \frac{\mathbf{Q}\theta^t}{\|\mathbf{Q}\theta^t\|_2}. \tag{8.46}$$

(a) Prove that there is a large set of initial vectors θ^0 for which the sequence $\{\theta^t\}_{t=0}^{\infty}$ converges to θ^*.

(b) Give a "bad" initialization for which this convergence does not take place.

(c) Based on part (b), specify a procedure to compute the second largest eigenvector, assuming it is also unique.

Exercise 8.6 (PCA for Gaussian mixture models)**.** Recall the Gaussian mixture model introduced in Example 8.3, and the PCA-based estimator $\widehat{\theta}$ for the mean vector θ^*.

(a) Prove that if the sample size is lower bounded as $n > c_1(\sigma^2 + \|\theta^*\|_2^2)d$ for a sufficiently large constant c_1, this estimator satisfies a bound of the form

$$\|\widehat{\theta} - \theta^*\|_2 \leq c_2 \sigma \sqrt{\frac{d}{n}}$$

with high probability.

(b) Explain how to use your estimator to build a classification rule—that is, a mapping $x \mapsto \psi(x) \in \{-1, +1\}$, where the binary labels code whether sample x has mean $-\theta^*$ or $+\theta^*$.

(c) Does your method still work if the shared covariance matrix is *not* a multiple of the identity?

Exercise 8.7 (PCA for uniform mixture models)**.** This exercise is a continuation of Exercise 8.6. For any scalar $a > 0$, let $\varphi_d(\cdot; a)$ denote the density of a uniform distribution on the ball $[-a, +a]^d$. Suppose that we observe n i.i.d. samples $\{x_i\}_{i=1}^n$ from the mixture distribution

$$f(x) = \alpha \varphi_d(x; a) + (1 - \alpha) \varphi_d(x; b) \tag{8.47}$$

where $a > b > 0$, and $\alpha \in (0, 1)$ is a mixing weight. Suppose that our goal is to build a classification rule $x \mapsto \{-1, +1\}$, coding the a-uniform and b-uniform components respectively.

(a) Is ordinary PCA applied to the samples $\{x_i\}_{i=1}^n$ useful in distinguishing between the two classes?

(b) Suggest how PCA can be applied to a simple transformation of the samples so as to yield an interesting answer.

Exercise 8.8 (PCA for retrieval from absolute values)**.** Suppose that our goal is to estimate an unknown vector $\theta^* \in \mathbb{R}^d$ based on n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ of the form $y_i = \big|\langle x_i, \, \theta^* \rangle\big|$. This model is a real-valued idealization of the problem of phase retrieval, to be discussed at more length in Chapter 10.

(a) Suggest a PCA-based method for estimating θ^*.

(b) Suppose that each $x_i \in \mathbb{R}^d$ is drawn i.i.d. from a standard normal distribution. Prove that as long as $n > c_1 d$ for a sufficiently large constant c_1, your estimator from part (a) satisfies a bound of the form $\|\widehat{\theta} - \theta^*\|_2 \leq c_2(1 + \|\theta^*\|_2) \sqrt{\frac{d}{n}}$ with high probability.

Exercise 8.9 (Semidefinite relaxation of sparse PCA)**.** Recall the nonconvex problem (8.27), also known as the `SCOTLASS` method. In this exercise, we derive a convex relaxation of the objective, due to d'Aspremont et al. [dEJL07].

(a) Show that the nonconvex problem (8.27) is equivalent to the optimization problem

$$\max_{\theta \in \mathcal{S}_+^{d \times d}} \operatorname{trace}(\widehat{\mathbf{\Sigma}} \mathbf{\Theta}) \qquad \text{such that } \operatorname{trace}(\mathbf{\Theta}) = 1, \ \textstyle\sum_{j,k=1}^d |\Theta_{jk}| \leq R^2, \text{ and } \operatorname{rank}(\mathbf{\Theta}) = 1,$$

where $\mathcal{S}_+^{d \times d}$ denotes the cone of symmetric, positive semidefinite matrices.

(b) Dropping the rank constraint yields the convex program

$$\max_{\theta \in \mathcal{S}_+^{d \times d}} \operatorname{trace}(\widehat{\mathbf{\Sigma}} \mathbf{\Theta}) \qquad \text{such that } \operatorname{trace}(\mathbf{\Theta}) = 1 \text{ and } \textstyle\sum_{j,k=1}^d |\Theta_{jk}| \leq R^2.$$

What happens when its optimum is achieved at a rank one matrix? Is the optimum always achieved at a rank one matrix?

Exercise 8.10 (Primal-dual witness for sparse PCA)**.** The SDP relaxation from Exercise 8.9(b) can be written in the equivalent Lagrangian form

$$\max_{\substack{\mathbf{\Theta} \in \mathcal{S}_+^{d \times d} \\ \operatorname{trace}(\mathbf{\Theta}) = 1}} \Big\{ \operatorname{trace}(\widehat{\mathbf{\Sigma}} \mathbf{\Theta}) - \lambda_n \sum_{j,k=1}^{d} |\Theta_{jk}| \Big\} \tag{8.48}$$

Suppose that there exists a vector $\widehat{\theta} \in \mathbb{R}^d$ and a matrix $\widehat{\mathbf{U}} \in \mathbb{R}^{d \times d}$ such that

$$\widehat{U}_{ij} = \begin{cases} \operatorname{sign}(\widehat{\theta}_j \widehat{\theta}_k) & \text{if } \widehat{\theta}_j \widehat{\theta}_k \neq 0, \text{ and} \\ \in [-1, 1] & \text{otherwise,} \end{cases} \tag{8.49a}$$

and moreover, such that

$$\operatorname{trace}\Big\{ \big(\widehat{\mathbf{\Sigma}} - \lambda_n \widehat{\mathbf{U}}\big) \mathbf{\Theta} \Big\} \leq \operatorname{trace}\Big\{ \big(\widehat{\mathbf{\Sigma}} - \lambda_n \widehat{\mathbf{U}}\big) \widehat{\theta}\, (\widehat{\theta})^T \Big\} \tag{8.49b}$$

for all $\mathbf{\Theta} \in \mathcal{S}_+^{d \times d}$ with $\operatorname{trace}(\mathbf{\Theta}) = 1$. Prove that the rank one matrix $\widehat{\mathbf{\Theta}} = \widehat{\theta} \otimes \widehat{\theta}$ is an optimal solution to the SDP relaxation (8.48).